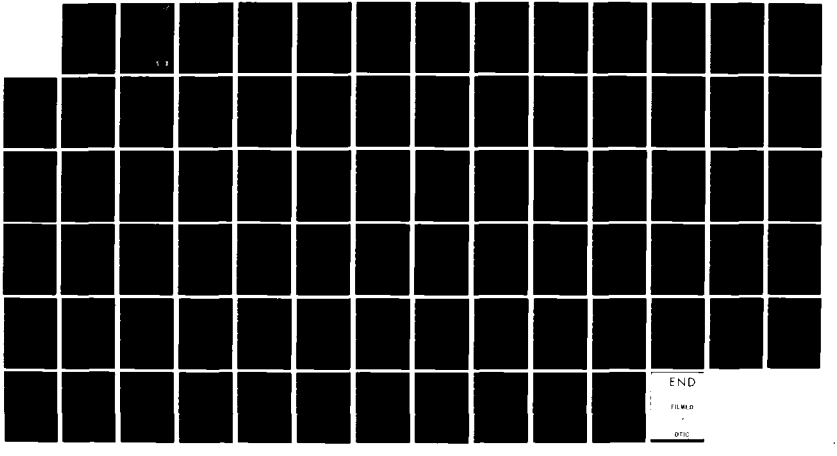
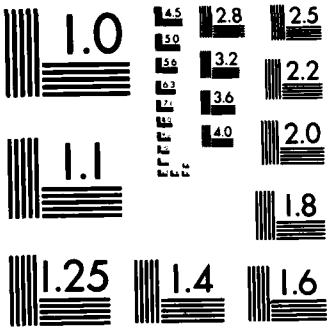
AD-A123 908 ESTIMATING OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND CORRELATION(U) CALIFORNIA UNIV BERKELEY DEPT OF STATISTICS L BREIMAN ET AL. JUL 82 TR-9 1/1

UNCLASSIFIED N00014-82-K-0054 F/G 12/1 NL





ESTIMATING OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND CORRELATION

BY
LEO BREIMAN[1] AND JEROME FRIEDMAN[2]

TECHNICAL REPORT NO. 9
JULY 1982

RESEARCH SUPPORTED IN PART
BY
[1]OFFICE OF NAVAL RESEARCH CONTRACT N00014-82-K-0054
[2]OFFICE OF NAVAL RESEARCH CONTRACT N00014-81-K-0340

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

Report 10 is ADA119802

ESTIMATING OPTIMAL TRANSFORMATIONS FOR MULTIPLE REGRESSION AND CORRELATION

Leo Breiman

Department of Statistics
University of California
Berkeley, California 94720

and

Jerome H. Friedman

Stanford Linear Accelerator Center
and
Department of Statistics
Stanford University
Stanford, California 94305

Abstract

In regression analysis the response variable Y and the predictor variables $X_1,\ldots,X_p$ are often replaced by functions $\theta(Y)$ and $\phi_1(X_1),\ldots,\phi_p(X_p)$. We discuss a procedure for estimating those functions θ^* and $\phi_1^*,\ldots,\phi_p^*$ that minimize

$$e^2 = \frac{E\{[\theta(Y) - \sum_{j=1}^{p}\phi_j(X_j)]^2\}}{Var[\theta(Y)]}$$

given only a sample $\{(y_k,x_{k1},\ldots,x_{kp}),\ 1\le k\le N\}$ and making minimal assumptions concerning the data distribution or the form of the solution functions. For the bivariate case, $p=1$, θ^* and ϕ^* satisfy $\rho^* = \rho(\theta^*,\phi^*) = \max_{\theta\phi}\rho[\theta(Y),\phi(X)]$ where ρ is the product moment correlation coefficient and ρ^* is the maximal correlation between X and Y. Our procedure thus also provides a method for estimating the maximal correlation between two variables.

*Work supported by Office of Naval Research under contracts N00014-82-K-0054 and N00014-81-K-0340.

KEY WORDS: Transformations, regression, correlation, smoothing

1. Introduction

Nonlinear transformation of variables is a commonly used practice in regression problems. Two common goals are stabilization of error variance and symmetrization/normalization of error distribution. A more comprehensive goal, and the one we adopt, is to find those transformations that produce the best fitting additive model. Knowledge of such transformations aid in the interpretation and understanding of the relationship between the response and predictors.

Let $Y,X_1,\ldots,X_p$ be random variables with Y the response and $X_1,\ldots,X_p$ the predictors. Let $\theta(Y),\phi_1(X_1),\ldots,\phi_p(X_p)$ be arbitrary measurable functions of the corresponding random variables. The fraction of variance not explained (e^2) by a regression of $\theta(Y)$ on $\sum_{i=1}^{p}\phi_i(X_i)$ is

$$e^2(\theta,\phi_1,\ldots,\phi_p) = \frac{E\{[\theta(Y) - \sum_{i=1}^{p}\phi_i(X_i)]^2\}}{Var[\theta(Y)]} \quad . \tag{1.1}$$

Then define *optimal transformations* as functions $\theta^*,\phi_1^*,\ldots,\phi_p^*$ that minimize (1.1): i.e.

$$e^2(\theta^*,\phi_1^*,\ldots,\phi_p^*) = \min_{\theta,\phi_1,\ldots,\phi_p} e^2(\theta,\phi_1,\ldots,\phi_p) \quad . \tag{1.2}$$

We show in Section 5 that optimal transformations exist and satisfy a complex system of integral equations. The heart of our approach is that there is a simple iterative algorithm using only bivariate conditional expectations which converges to an optimal solution. When the conditional expectations are estimated from a finite data set, then use of the algorithm results in estimates of the optimal transformations.

This method has some powerful characteristics. It can be applied in situations where the response and/or the predictors involve arbitrary mixtures of continuous ordered variables and categorical variables (ordered or unordered). The functions $\theta,\phi_1,\ldots,\phi_p$ are real valued. If the original variable is categorical, the application of θ or ϕ_i assigns a real valued score to each of its categorical values.

The procedure is nonparametric. The optimal transformation estimates are based solely on the data sample $\{(y_k,x_{k1},\ldots,x_{kp}),\ 1 \leq k \leq N\}$ with minimal assumptions concerning the data distribution and the form of the optimal transformations. The principal distributional assumption is that the data are i.i.d. In particular, we do not require the transformation functions to be from a particular parameterized family or even monotone. (We illustrate below situations where the optimal transformations are not monotone.)

For the bivariate case, $p=1$, the optimal transformations $\theta^*(Y)$, $\phi^*(X)$ satisfy

$$\rho^*(X,Y) = \rho(\theta^*,\phi^*) = \max_{\theta,\phi} \rho[\theta(Y),\phi(X)] \tag{1.3}$$

where ρ is the product moment correlation coefficient. The quantity $\rho^*(X,Y)$ is known as the *maximal correlation* between X and Y, and is used as a general measure of dependence (Gebelern [1947]; see also Renyi [1959] and Sarmanov [1958A,B]). The maximal correlation has the following properties (Renyi [1959]):

(a) $0 \leq \rho^*(X,Y) \leq 1$

(b) $\rho^*(X,Y) = 0$ if and only if X and Y are independent

(c) if there exists a relation of the form $u(X) = v(Y)$ where u and v are Borel-measurable functions with $\text{var}[u(X)] > 0$, then $\rho^*(X,Y) = 1$.

Therefore, in the bivariate case our procedure can also be regarded as a method for estimating the maximal correlation between two variables, providing as a by-product estimates of the functions θ^*, ϕ^* that achieve the maximum.

In the next section, we describe our procedure for finding optimal transformations using algorithmic notation, deferring mathematical justifications to sections 5 and 6. We next illustrate the procedure in Section 3 by applying it to a number of simulated data sets where the optimal transformations are known. The estimates are surprisingly good. Our algorithm is also applied to the Boston housing data of Harrison and Rubinfeld [1978] as listed in Belsey, Kuh and Welsch [1980]. The transformations found by the algorithm generally differ from those applied in the original analysis. (A FORTRAN implementation of our algorithm is listed in Appendix 2). Section 4 presents a general discussion and relates this procedure to other empirical methods for finding transformations.

Sections 5, 6, and Appendix 1 provide some theoretical framework for the algorithm. In Section 5, under weak conditions on the joint distribution of $Y,X_1,\ldots,X_p$, it is shown that optimal transformations exist and are generally unique up to a change of sign. The optimal transformations are characterized as the eigenfunctions of a set of linear integral equations whose kernels involve bivariate distributions. We then show that our procedure converges to optimal transformations.

Section 6 discusses the algorithm as applied to finite data sets. The results are dependent on the type of data smooth employed to estimate

the bivariate conditional expectations. Convergence of the algorithm is proven only for a very restricted class of data smooths. However, in over a thousand applications of the algorithm on a variety of data sets using three different types of data smoothers only one (very contrived) instance of non convergence has been found.

Section 6 also contains proof of a consistency result. Under fairly general conditions, as the sample size increases the finite data transformations converge in a "weak" sense to the distributional space optimal transformations. Finally, Appendix 1 contains a brief discussion of the needed consistency properties of bivariate smooths.

This paper is laid out in two distinct parts. Sections 1-4 give a fairly non-technical overview of the method and discuss its application to data. Sections 5 and 6 are, of necessity, more technical, presenting the theoretical foundation for the procedure.

2. The Algorithm

Our procedure for finding $\theta^*,\phi_1^*,\ldots,\phi_p^*$ is iterative. Assume a known distribution for the variables $Y,X_1,\ldots,X_p$. Without loss of generality, let $\mathrm{var}[\theta(Y)] = 1$, and assume that all functions have expectation zero.

To illustrate, we first look at the bivariate case;

$$e^2(\theta,\phi) = E[\theta(Y)-\phi(X)]^2 \tag{2.1}$$

Consider the minimization of (2.1) with respect to $\theta(Y)$ for a given function $\phi(X)$. The solution is

$$\theta'(Y) = E[\phi(X)|Y]/\|E[\phi(X)|Y]\| \tag{2.2}$$

with $\|\cdot\| \equiv [E(\cdot)^2]^{1/2}$. Next, consider the minimization of (2.1) with respect to $\phi(X)$ for a given $\theta(Y)$. The solution is

$$\phi'(X) = E[\theta(Y)|X] \ . \tag{2.3}$$

Equations (2.2) and (2.3) form the basis of an iterative optimization procedure involving alternating conditional expectations (ACE):

```
BASIC ACE ALGORITHM
   set θ(Y) = Y/‖Y‖;
   ITERATE UNTIL e²(θ,φ) fails to decrease:
      φ'(X) = E[θ(Y)|X];
      replace φ(X) with φ'(X);
      θ'(Y) = E[φ(X)|Y]/‖E[φ(X)|Y]‖;
      replace θ(Y) with θ'(Y);
   END ITERATION LOOP;
   θ and φ are the solutions θ* and φ*;
END ALGORTHM;
```

This algorithm decreases (2.1) at each step by alternatingly minimizing with respect to one function holding the other fixed at its previous evaluation. Each iteration (execution of the iteration loop) performs one pair of these single function minimizations. The process begins with an initial guess for one of the functions ($\theta = Y/\|Y\|$ above) and ends when a complete iteration pass fails to decrease e^2 (2.1). In Section 5, we prove that the algorithm converges to optimal transformations θ^*, ϕ^*.

Now consider the more general case of multiple predictors $X_1,\ldots,X_p$. We proceed in direct analogy with the basic ACE algorithm; we minimize

$$e^2(\theta,\phi_1,\ldots,\phi_p) = E[\theta(Y) - \sum_{j=1}^{p}\phi_j(X_j)]^2 \quad , \tag{2.4}$$

holding $E\theta^2 = 1$, $E\theta = E\phi_1 = \cdots = E\phi_p = 0$, through a series of single function minimizations involving bivariate conditional expectations. For a given set of functions $\phi_1(X_1),\ldots,\phi_p(X_p)$, minimization of (2.4) with respect to $\theta(Y)$ yields

$$\theta'(Y) = E[\sum_{i=1}^{p}\phi_i(X_i)|Y]/\|E[\sum_{i=1}^{p}\phi_i(X_i)|Y]\| \tag{2.5}$$

The next step is to minimize (2.4) with respect to $\phi_1(X_1),\ldots,\phi_p(X_p)$ given $\theta(Y)$. This is obtained through another iterative algorithm. Consider the minimization of (2.4) with respect to a single function $\phi_k(X_k)$ for given $\theta(Y)$ and a given set $\phi_1,\ldots,\phi_{k-1},\phi_{k+1},\ldots,\phi_p$. The solution is

$$\phi_k'(X_k) = E[\theta(Y) - \sum_{i\neq k}\phi_i(X_i)|X_k] \ . \tag{2.6}$$

The corresponding iterative algorithm is then:

set $\phi_1(X_1),\ldots,\phi_p(X_p) = 0$;
ITERATE UNTIL $e^2(\theta,\phi_1,\ldots,\phi_p)$ fails to decrease;
  FOR k = 1 TO p DO;
    $\phi_k'(X_k) = E[\theta(Y) - \sum_{i\neq k} \phi_i(X_i)|X_k]$;
    replace $\phi_k(X_k)$ with $\phi_k'(X_k)$;
  END FOR LOOP;
END ITERATION LOOP;
$\phi_1,\ldots,\phi_p$ are the solution functions;

Each iteration of the inner FOR loop minimizes e^2 (2.4) with respect to the function $\phi_k(X_k)$, k = 1,...,p with all other functions fixed at their previous evaluations (execution of the FOR loop). The outer loop is iterated until one complete pass over the predictor variables (inner FOR loop) fails to decrease e^2 (2.4).

Substituting this procedure for the corresponding single function optimization in the bivariate ACE algorithm gives rise to the full ACE algorithm for minimizing e^2 (2.4):

ACE ALGORITHM:

  set $\theta(Y) = Y/\|Y\|$ and $\phi_1(X_1),\ldots,\phi_p(X_p) = 0$;
  ITERATE UNTIL $e^2(\theta,\phi_1,\ldots,\phi_p)$ fails to decrease;
    ITERATE UNTIL $e^2(\theta,\phi_1,\ldots,\phi_p)$ fails to decrease;
      FOR k = 1 TO p DO:
        $\phi_k'(X_k) = E[\theta(Y) - \sum_{i\neq k} \phi_i(X_i)|X_k]$;
        replace $\phi_k(X_k)$ with $\phi_k'(X_k)$;
      END FOR LOOP;
    END INNER ITERATION LOOP;
    $\theta'(Y) = E[\sum_{i=1}^{p} \phi_i(X_i)|Y]/\|E[\sum_{i=1}^{p} \phi_i(X_i)|Y]\|$
    replace $\theta(Y)$ with $\theta'(Y)$;
  END OUTER ITERATION LOOP;
  $\theta,\phi_1,\ldots,\phi_p$ are the solutions $\theta^*,\phi_1^*,\ldots,\phi_p^*$;
END ACE ALGORTHM;

In Section 5, we prove that the ACE algorithm converges to optimal transformations.

3. Applications

In the previous section, the ACE algorithm was developed in the context of known distributions. In practice, data distributions are seldom known. Instead, one has a data set $\{(y_k, x_{k1}, \ldots, x_{kp}),\ 1 \le k \le N\}$ that is presumed to be a sample from $Y, X_1, \ldots, X_p$. The goal is to estimate the optimal transformation functions $\theta(Y), \phi_1(X_1), \ldots, \phi_p(X_p)$ from the data. This can be accomplished by applying the ACE algorithm to the data wi' the quantity e^2, $\| \ \|$, and the conditional expectations replaced by suitable estimates. The resulting functions $\hat{\theta}^*, \hat{\phi}_1^*, \ldots, \hat{\phi}_p^*$ are then taken as estimates of the corresponding optimal transformations.

The estimate for e^2 is the usual mean squared error for regression,

$$e^2(\theta, \phi_1, \ldots, \phi_p) = \frac{1}{N} \sum_{k=1}^{N} [\theta(y_k) - \sum_{j=1}^{p} \phi_j(x_{kj})]^2$$

If $g(y, x_1, \ldots, x_p)$ is a function defined for all data values, then $\|g\|^2$ is replaced by

$$\|g^2\|_N = \frac{1}{N} \sum_{k=1}^{N} g^2(y_k, x_{k1}, \ldots, x_{kp}) \ .$$

For the case of categorical variables, the conditional expectation estimates are straightforward:

$$\hat{E}[A|Z = z] = \sum_{z_j = z} A_j \Big/ \sum_{z_j = z} 1$$

where A is a real valued quantity and the sums are over the subset of observations having (categorical) value $Z = z$. For variables that can assume many ordered values, the estimation is based on smoothing techniques. Such procedures have been the subject of considerable study (see, for example, Gasser and Rosenblatt [1979], Cleveland [1979], Craven and

Wahba [1979]). Since the smoother is repeatedly applied in the algorithm, high speed is desirable, as well as adaptability to local curvature. We use a smoother employing local linear fits with varying window width determined by local cross-validation ("super smoother", Friedman and Stuetzle [1982]).

The algorithm evaluates $\hat{\theta}^*,\hat{\phi}_1^*,\ldots,\hat{\phi}_p^*$ at all the corresponding data values, i.e. $\hat{\theta}^*(y)$ is evaluated at the set of data values $\{y_k\}$, $k=1,\ldots,N$. The simplest way to understand the shape of the transformations is by means of a plot of the function versus the corresponding data values, that is, through the plots of $\hat{\theta}^*(y_k)$ versus y_k and $\hat{\phi}_1^*,\ldots,\hat{\phi}_p^*$ versus the data values of $x_1,\ldots,x_p$ respectively.

In this section, we illustrate the ACE procedure by applying it to various data sets. (A FORTRAN subroutine implementing our procedure is listed in Appendix 2.) In order to evaluate performance on finite samples, the procedure is first applied to simulated data for which the optimal transformations are known. We then apply it to the Boston housing data of Harrison and Rubinfeld [1978] as listed in Belsey, Kuh and Welsch [1980], contrasting the ACE transformations with those used in the original analysis.

Our first example consists of 200 bivariate observations $\{(y_k,x_k), 1 \leq k \leq 200\}$ generated from the model

$$y_k = \exp[\sin(x_k) + e_k/2]$$

with the x_k sampled from a uniform distribution $U(0,2\pi)$ and the e_k drawn independently of the x_k from a standard normal distribution $N(0,1)$. Figure 1a shows a scatterplot of these data. Figures 1b-1d show the results of applying the ACE algorithm to the data. The estimated

optimal transformation $\hat{\theta}^*(y)$ is shown in the plot 1b of $\hat{\theta}^*(y_k)$ versus y_k, $1 \leq k \leq 200$. Figure 1c is a plot of $\hat{\phi}^*(x_k)$ versus x_k. These plots clearly suggest the transformations $\theta(y) = \log(y)$ and $\phi(x) = \sin(x)$ which are optimal for the parent distribution. Figure 1d is a plot of $\hat{\theta}^*(y_k)$ versus $\hat{\phi}^*(x_k)$. This plot indicates a more linear relation between the transformed variables than that between the untransformed ones.

The next issue we address is how much the algorithm overfits the data due to the repeated smoothings, resulting in inflated estimates of the maximal correlation ρ^* and of $R^{*2} = 1 - e^{*2}$. The answer, on the simulated data sets we have generated, is surprizing little.

To illustrate this, we contrast two estimates of ρ^* and R^{*2} using the above model. The known optimal transformations are $\theta(Y) = \log Y$, $\phi(X) = \sin X$. Therefore, we define the *direct* estimate for ρ^* given any data set generated as above by

$$\hat{\rho}^* = \frac{1}{N} \sum_{k=1}^{N} (\log y_k - \overline{\log y})(\sin x_k - \overline{\sin x})$$

and $\hat{R}^{*2} = \hat{\rho}^{*2}$. The ACE algorithm produces the estimates

$$\hat{\rho}^* = \frac{1}{N} \sum_{k=1}^{N} \hat{\theta}^*(y_k)\hat{\phi}^*(x_k)$$

and $\hat{R}^{*2} = 1 - \hat{e}^{*2} = \hat{\rho}^{*2}$. In this model $\rho^* = .8165$ and $R^{*2} = .6667$.

For 100 data sets, each of size 200, generated from the above model, the means and standard deviations of the ρ^* estimates are

	mean	s.d.
ρ^* direct	.814	.022
ACE	.808	.031

Figure 1a

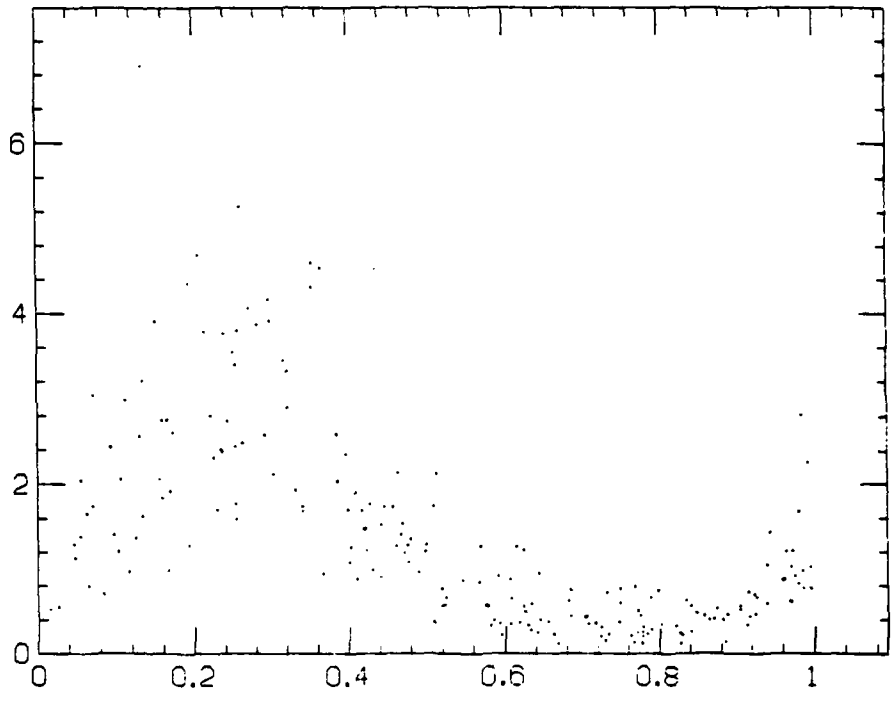


Figure 1b

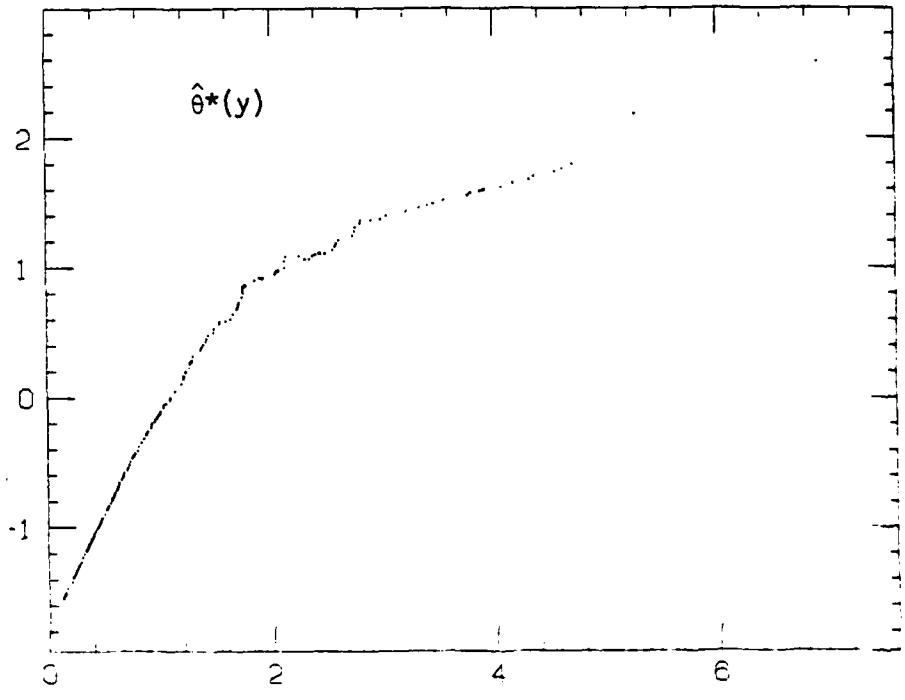


Figure 1c

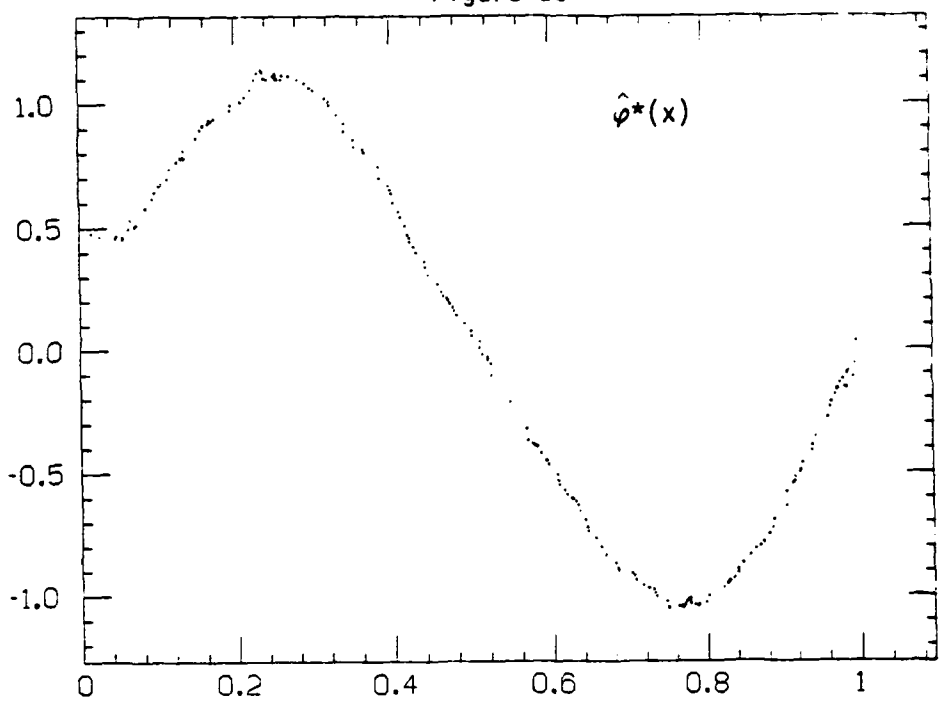
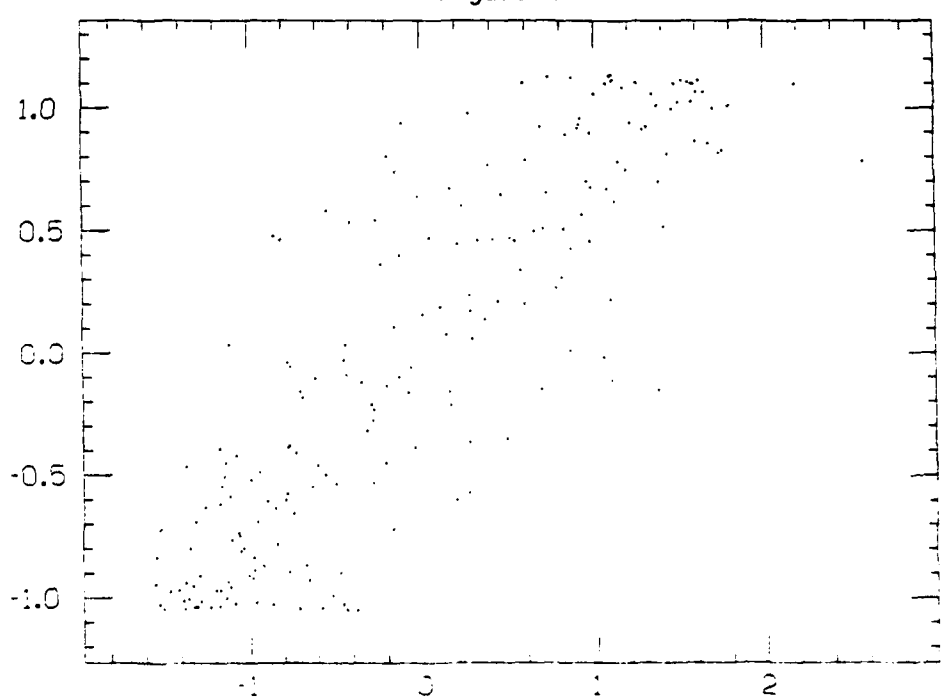


Figure 1d



The means and standard deviations of the R^{*2} estimates are

	mean	s.d.
R^{*2} direct	.664	.031
ACE	.654	.050

We also computed the differences $\hat{\rho}^* - \hat{\hat{\rho}}^*$ and $\hat{R}^{*2} - \hat{\hat{R}}^{*2}$ for the 100 data sets. The means and standard deviations are

	mean	s.d.
$\hat{\rho}^* - \hat{\hat{\rho}}^*$	-.006	.015
$\hat{R}^{*2} - \hat{\hat{R}}^{*2}$	-.010	.024

The above experiment was duplicated for smaller sample size $N = 100$. In this case we obtain

	mean	s.d.
$\hat{\rho}^* - \hat{\hat{\rho}}^*$	-.005	.027
$\hat{R}^{*2} - \hat{\hat{R}}^{*2}$	-.007	.044

Our next example consists of a sample of 200 triples $\{(y_k, x_{k1}, x_{k2}), 1 \leq k \leq 200\}$ drawn from the model $Y = X_1 X_2$ with X_1 and X_2 generated independently from a uniform distribution $U(-1,1)$. Note that $\theta(Y) = \log(Y)$ and $\phi_j(X_j) = \log X_j$ $(j = 1,2)$ cannot be solutions here since Y, X_1 and X_2 all assume negative values. Figure 2a shows a plot of $\hat{\theta}^*(y_k)$ versus y_k, while Figures 2b and 2c show corresponding plots of $\hat{\phi}_1^*(x_{k1})$ and $\hat{\phi}_2^*(x_{k2})$ $(1 \leq k \leq 200)$. All three solution transformation functions are seen to be double valued. The optimal transformations for this problem are $\theta^*(Y) = \log|Y|$ and $\phi_j^*(X_j) = \log|X_j|$ $(j = 1,2)$. The estimates clearly reflect this structure except near the origin where the smoother cannot reproduce the infinite discontinuity in the derivative.

Figure 2a

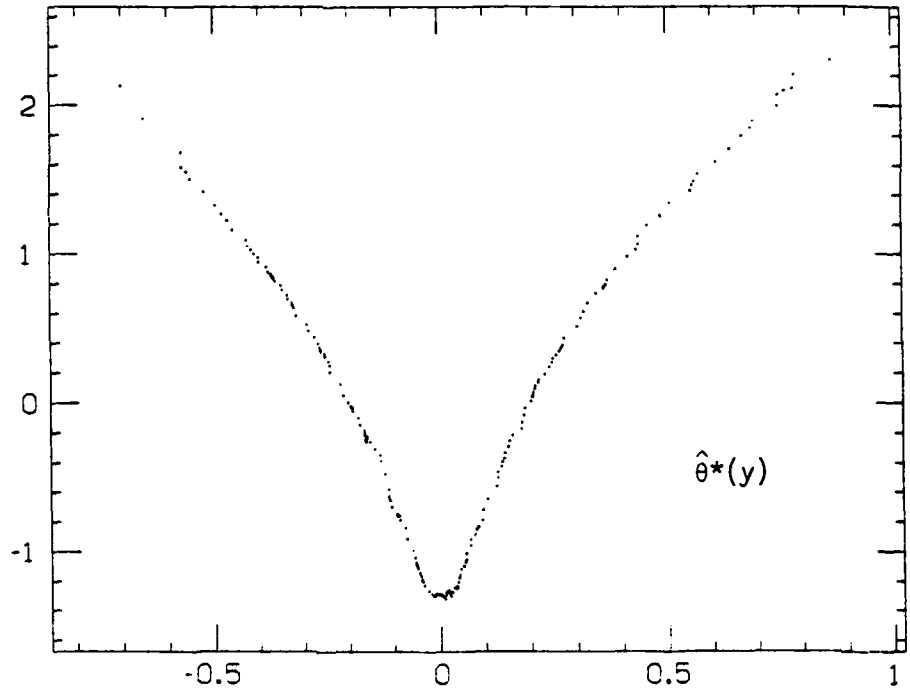


Figure 2b

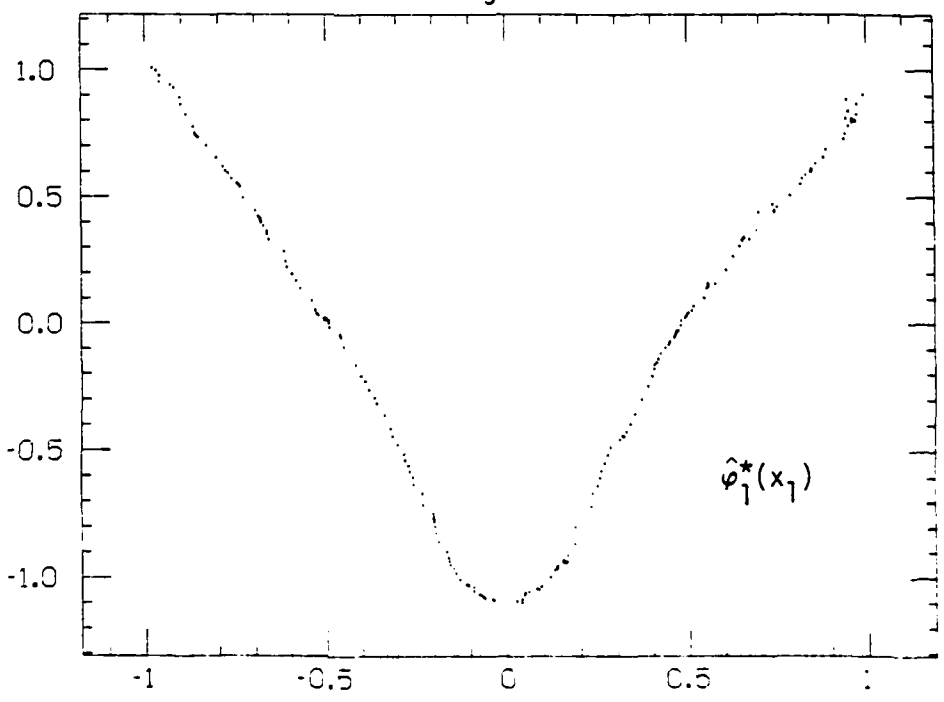
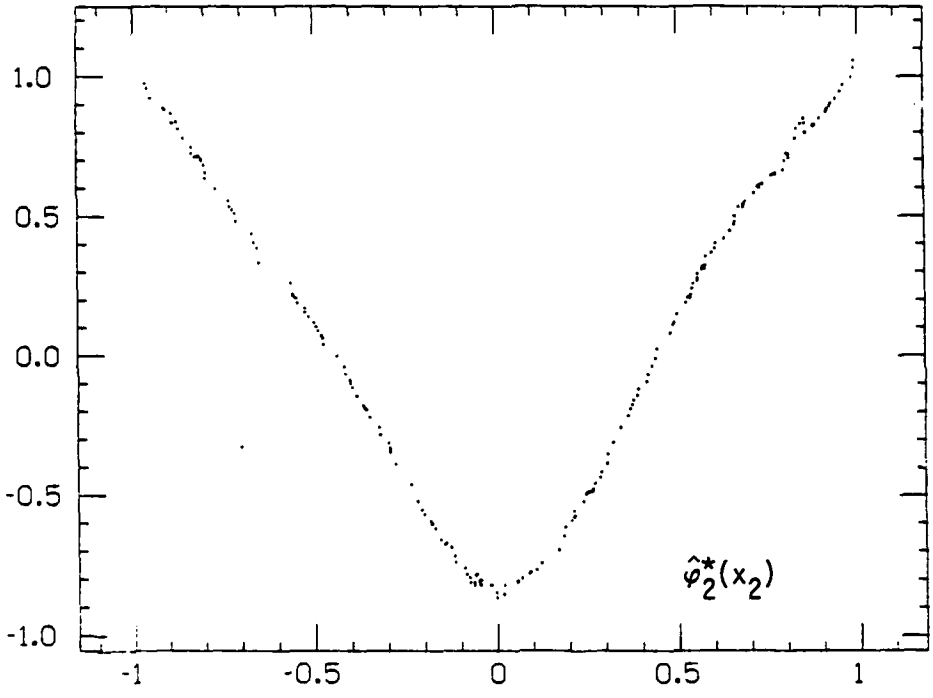


Figure 2c



For our final example, we apply the ACE algorithm to the Boston housing market data of Harrison and Rubinfeld [1978]. A complete listing of these data appear in Belsey, Kuh and Welsch [1980]. Harrison and Rubinfeld used these data to estimate marginal air pollution damages as revealed in the housing market. Central to their analysis was a housing value equation which relates the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Area, to air pollution (as reflected in concentration of nitrogen oxides) and to 12 other variables that are thought to effect housing prices. This equation was estimated by trying to determine the best fitting functional form of housing price on these 13 variables. By experimenting with a number of possible transformations of the 14 variables (response and 13 predictors), Harrison and Rubinfeld settled on an equation of the form

$$\begin{aligned}\log(MV) = \alpha_1 &+ \alpha_2(RM)^2 + \alpha_3\, AGE \\ &+ \alpha_4\, \log(DIS) + \alpha_5\, \log(RAD) + \alpha_6\, TAX \\ &+ \alpha_7\, PTRATIO + \alpha_8(B-0.63)^2 \\ &+ \alpha_9\, \log(LSTAT) + \alpha_{10}\, CRIM + \alpha_{11}\, ZN \\ &+ \alpha_{12}\, INDUS + \alpha_{13}\, CHAS + \alpha_{14}(NOX)^p + \epsilon\ .\end{aligned}$$

A brief description of each variable is given in Table 1. (For a more complete description, see Harrison and Rubinfeld [1978], Table IV.) The coefficients $\alpha_1,\ldots,\alpha_{14}$ were determined by a least squares fit to measurements of the 14 variables for the 506 census tracts. The best value for the exponent p was found to be 2.0, by a numerical optimization (grid search). This "basic equation" was used to generate estimates for the willingness to pay for and the marginal benefits of clean air.

TABLE 1

Variables Used in the Housing Value Equation
of Harrison and Rubinfeld (1978)

Variable	Definition
MV	Median value of owner-occupied homes
RM	Average number of rooms in owner units
AGE	Proportion of owner units built prior to 1940
DIS	Weighted distances to five employment centers in the Boston region
RAD	Index of accessibility to radial highways
TAX	Full property tax rate ($/$10,000)
PTRATIO	Pupil-teacher ratio by town school district
B	Black proportion of population
LSAT	Proportion of population that is lower status
CRIM	Crime rate by town
ZN	Proportion of town's residential land zoned for lots greater than 25,000 square feet
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy: = 1 if tract bounds the Charles River; = 0 if otherwise
NOX	Nitrogen oxide concentration in pphm

Harrison and Rubinfeld note that the results are highly sensitive to the particular specification of the form of the housing price equation.

We applied the ACE algorithm to the transformed measurements (using $p = 2$ for NOX) appearing in the basic equation. To the extent that these transformations are close to the optimal ones, the algorithm will produce results close to linear functions $\theta(Z) = \phi_1(Z) = \cdots = \phi_{13}(Z) = \alpha Z + \beta$. Departures from linearity indicate transformations that can improve the quality of the fit.

Figure 3a shows a plot of the solution response transformation $\hat{\theta}^*(\log y)$. This function is seen to have a positive curvature for central values of y, connecting two straight line segments of different slope on either side. This suggests that the logarithmic transformation may be too severe. Figure 3b shows the transformation $\hat{\theta}^*(y)$ resulting when the ACE algorithm is applied to the original *untransformed* census measurements. This indicates that, if anything, a very mild transformation, involving *positive* curvature, is most appropriate for the response variable.

Figures 3c-3o show the ACE transformations $\hat{\phi}_1^*, \ldots, \hat{\phi}_{13}^*$ for the (transformed) predictor variables. The standard deviation $\sigma(\hat{\phi}_j^*)$ is indicated in each graph. This provides a measure of how strongly each $\hat{\phi}_j^*(x_j)$ enters into the model for $\hat{\theta}^*(y)$. (Note that $\sigma(\hat{\theta}^*) = 1.$) The two terms that enter most strongly involve the number of rooms (Figure 3c) and the fraction of population that is of lower status (Figure 3j). The nearly linear shape of the latter transformation suggests that the original logarithmic transformation was appropriate for this variable. The transformation on the number of rooms variable is far from linear, however, indicating that a quadratic does not adequately capture its relationship

to housing value. For less than six rooms, housing value is roughly independent of room number, while for larger values there is a strong increasing linear dependence. Among the next three variables (in order of their contribution to the model), log DIS (Figure 3e), CRIM (Figure 3k), and INDUS (Figure 3m), only CRIM has a solution close to a straight line. The plots for the remaining variables indicate that several of them could, as well, benefit from transformations substantially different from those used to define the basic equation.

The marginal effect of $(NOX)^2$ on median home value, as captured by this model, can be investigated by studying $\hat{\phi}^*[(NOX)^2]$ in Figure 3o. This curve is a nonmonotonic function of NOX^2 not well approximated by a linear (or monotone) function. This makes it difficult to formulate a simple interpretation of the willingness to pay for clean air from these data. For low concentration values, housing prices seem to *increase* with increasing $(NOX)^2$, whereas for higher values this trend is substantially reversed.

Figure 3p shows a scatterplot of $\hat{\theta}^*(y_k)$ versus $\sum_{j=1}^{13} \hat{\phi}_j^*(x_{kj})$. This plot shows no evidence of additional structure not captured in the model

$$\hat{\theta}^*(y) = \sum_{j=1}^{13} \hat{\phi}_j^*(x_j) + e \ .$$

The $\hat{e}^{*2}$ resulting from the use of the ACE transformations was 0.11 as compared to the e^2 value of 0.20 produced by the Harrison and Rubinfeld [1978] transformations.

Figure 3a

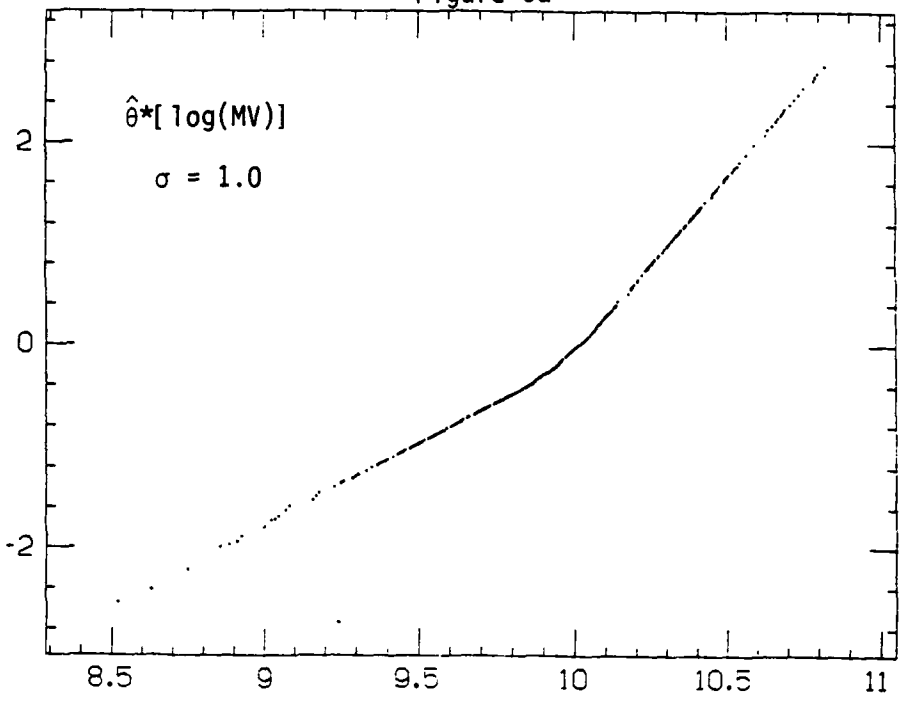


Figure 3b

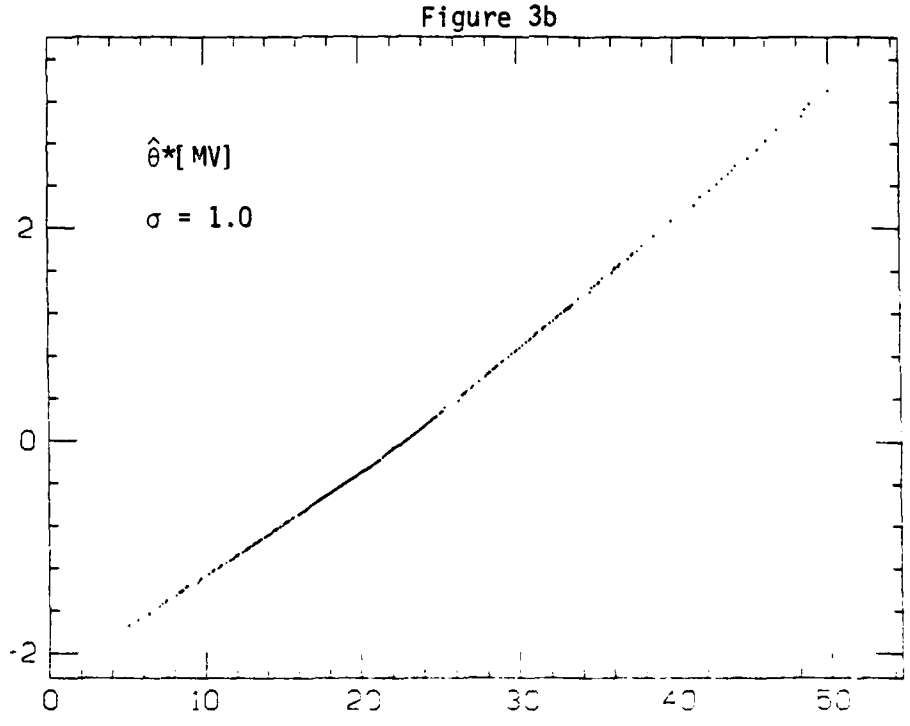


Figure 3c

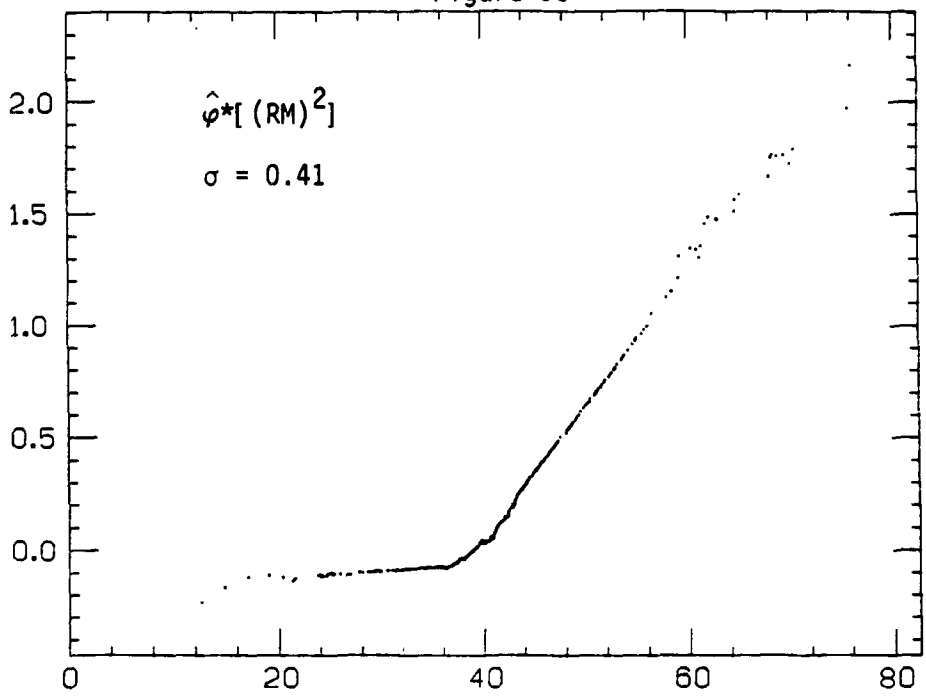


Figure 3d

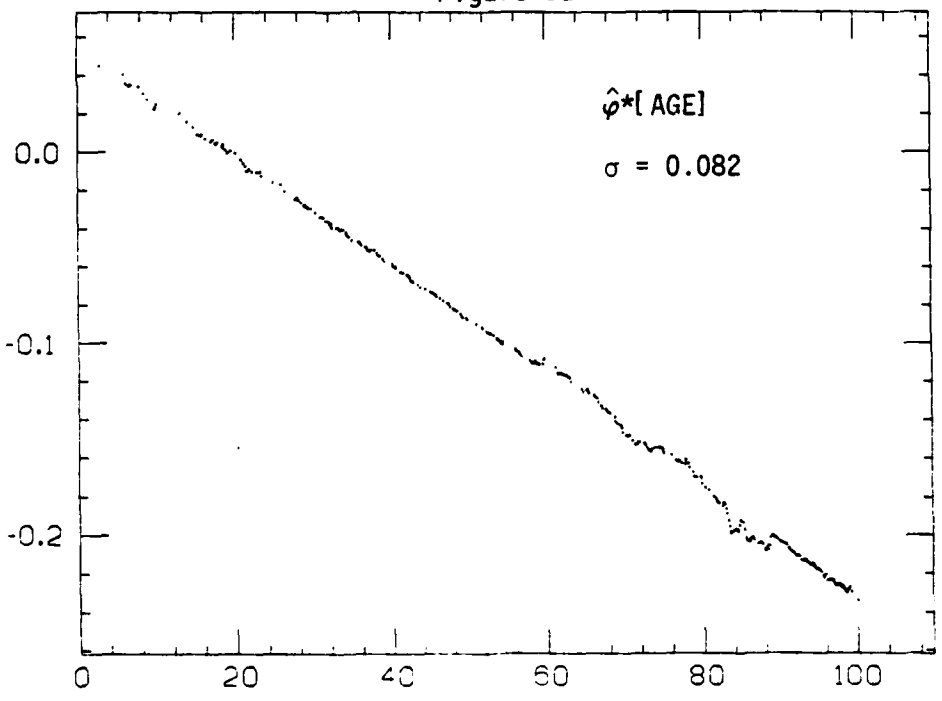


Figure 3e

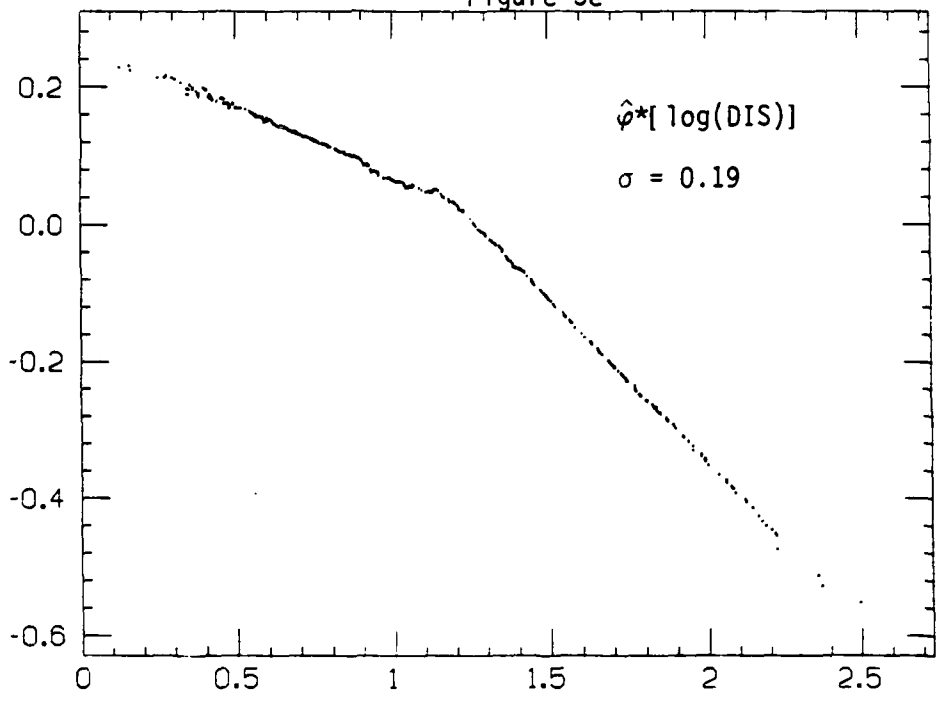


Figure 3f

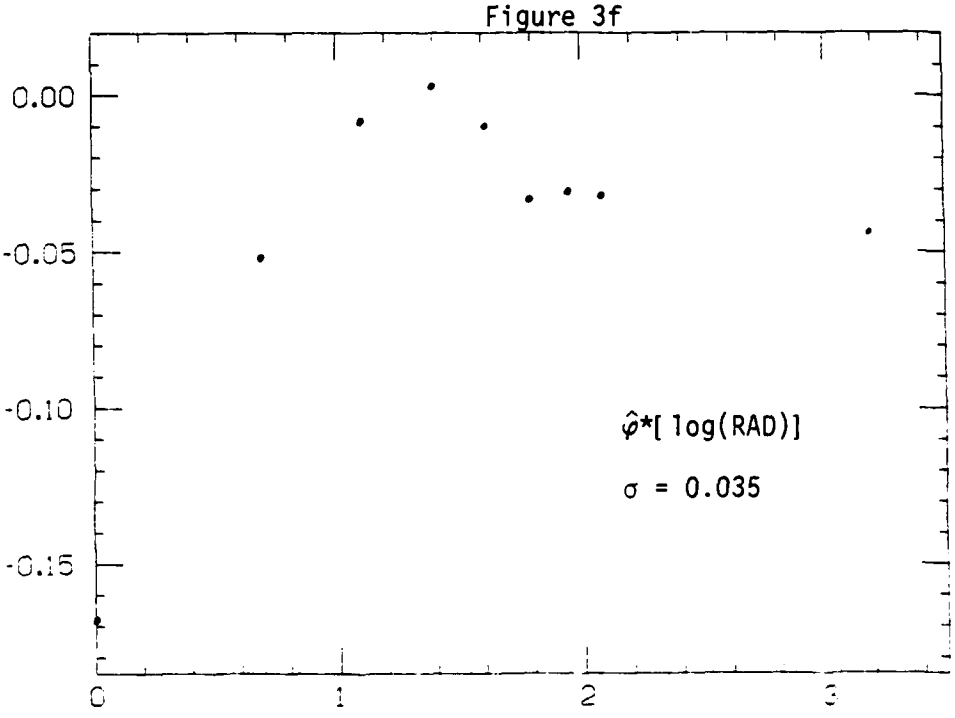


Figure 3g

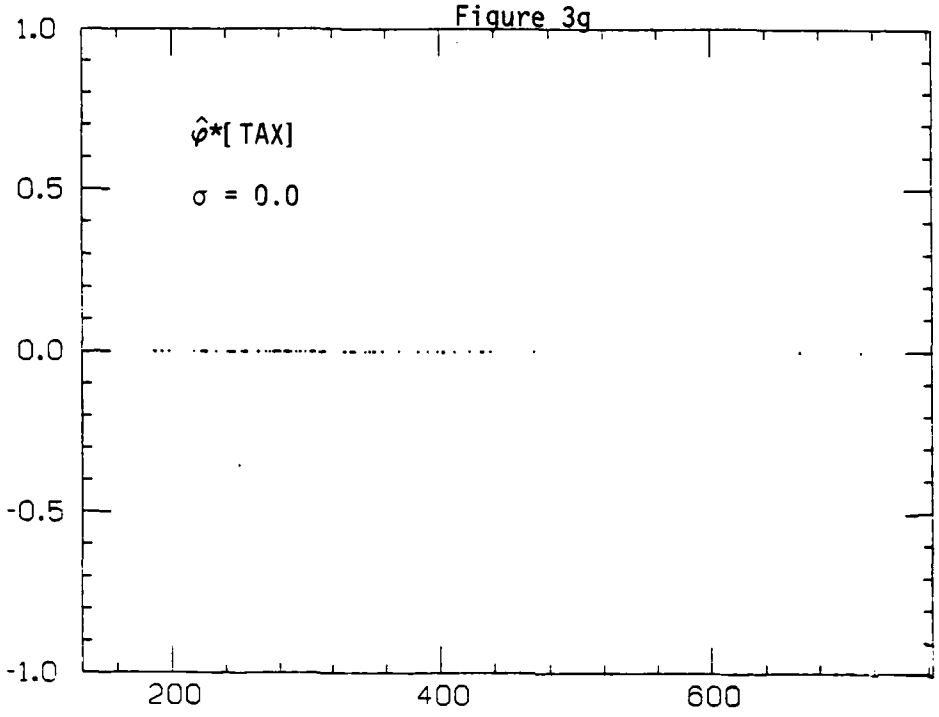


Figure 3h

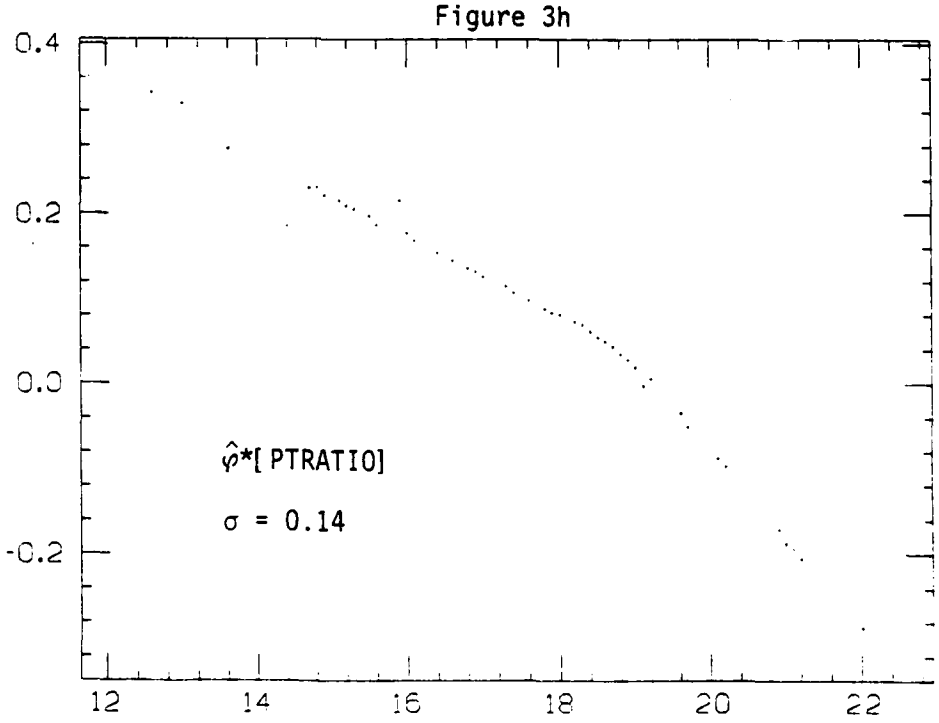


Figure 3i

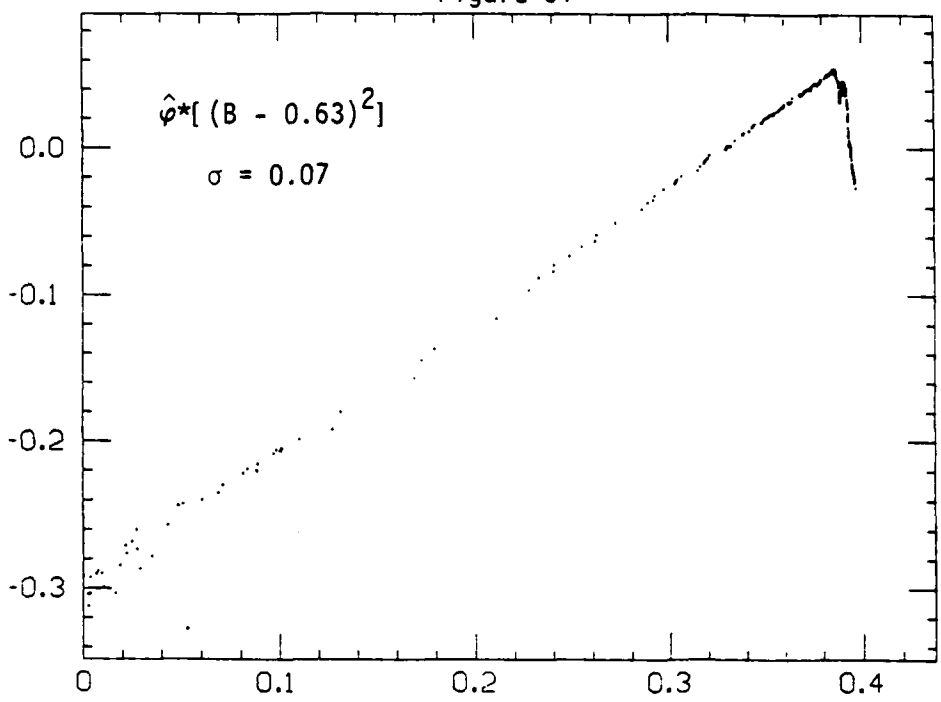


Figure 3j

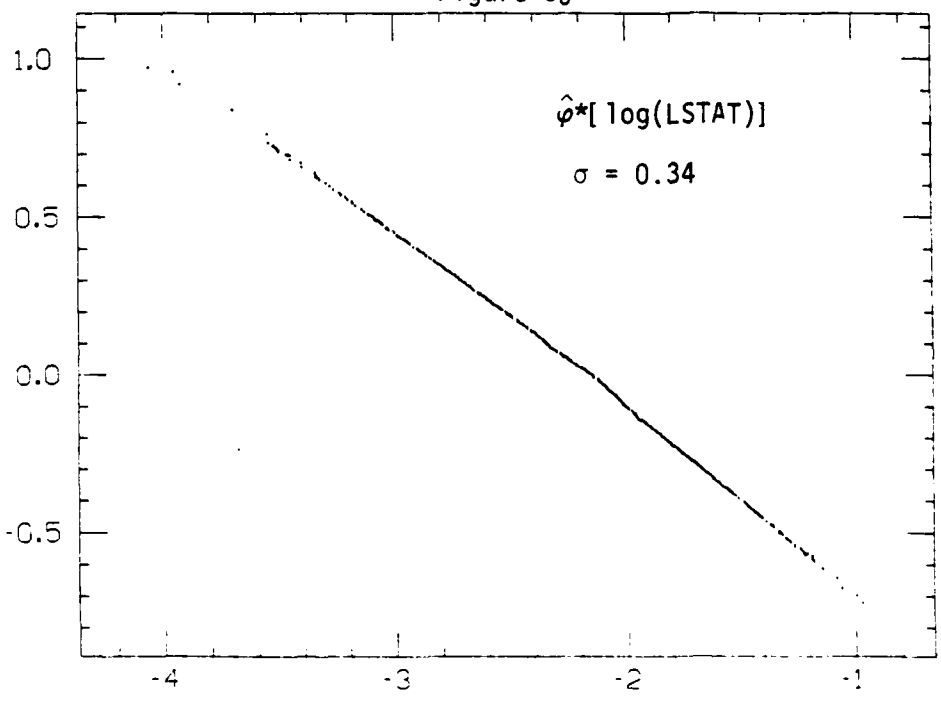


Figure 3k

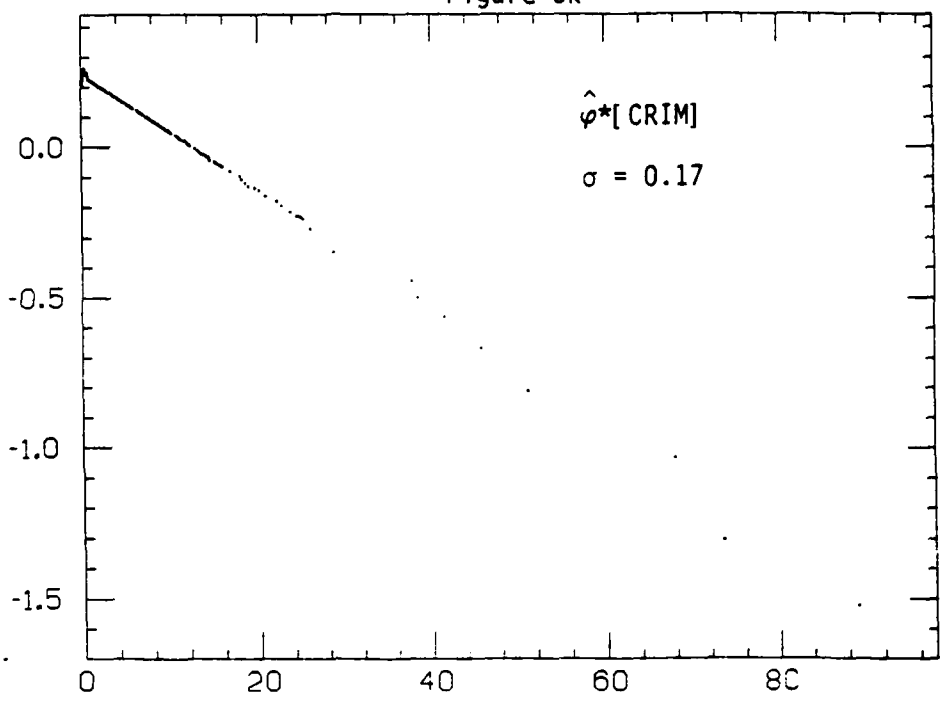


Figure 3ℓ

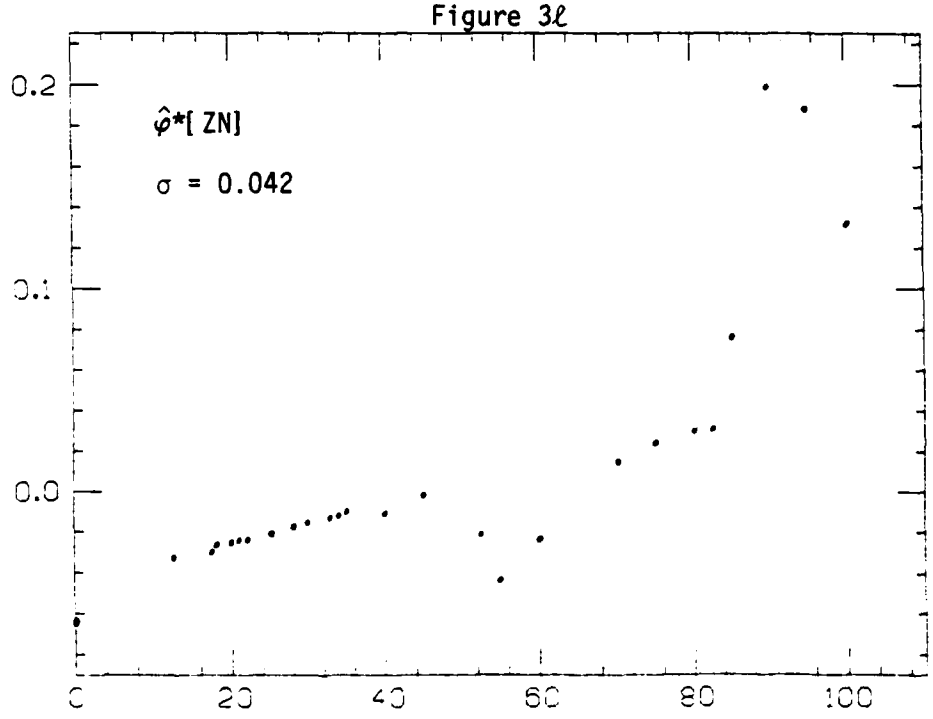


Figure 3m

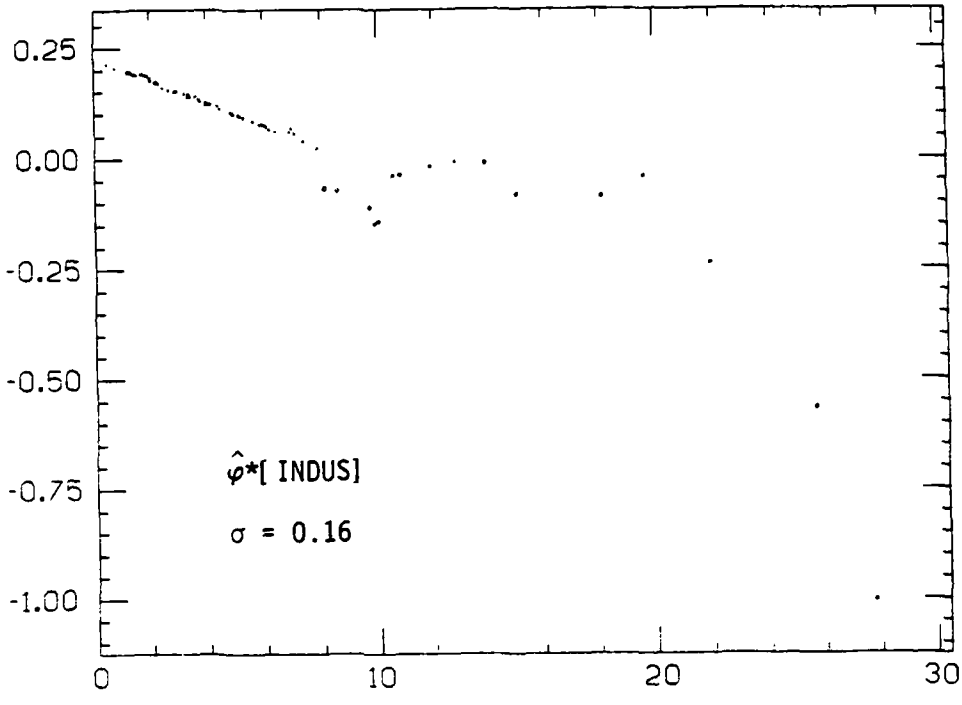


Figure 3n

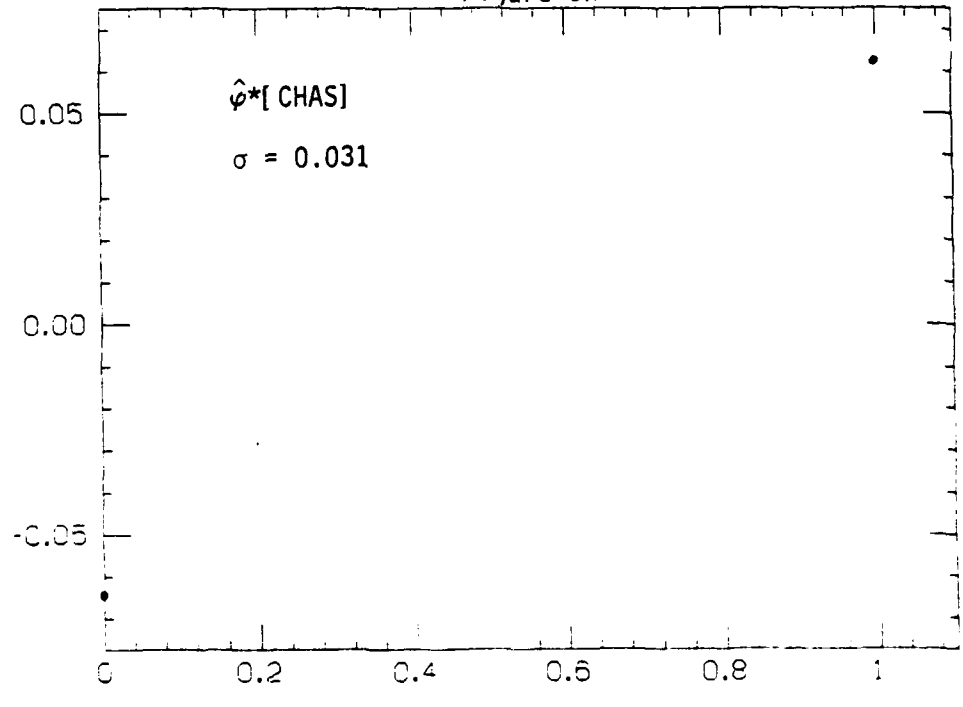


Figure 3o

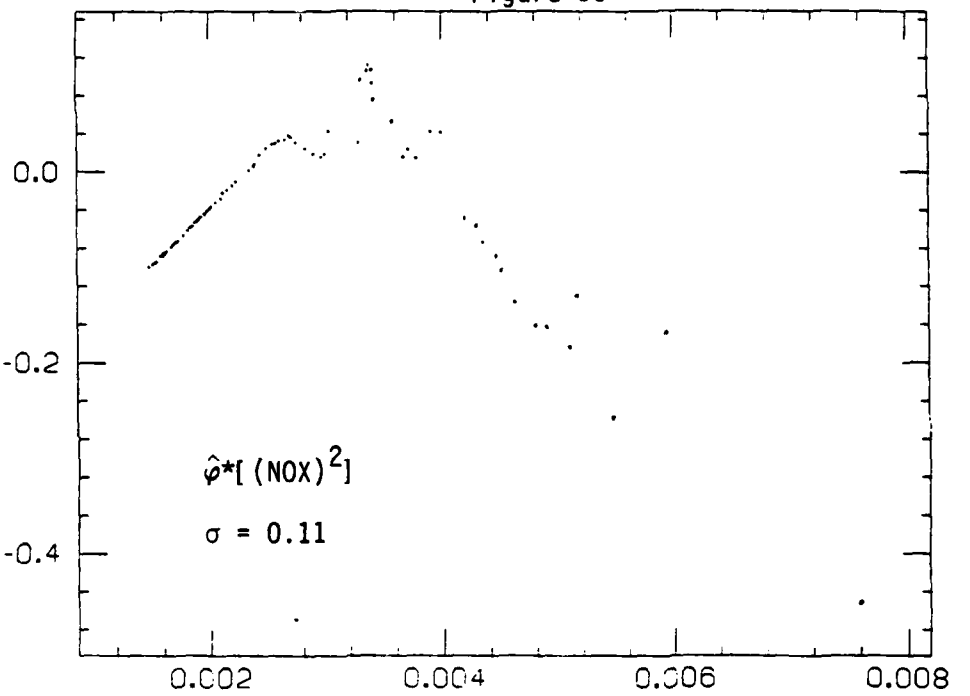
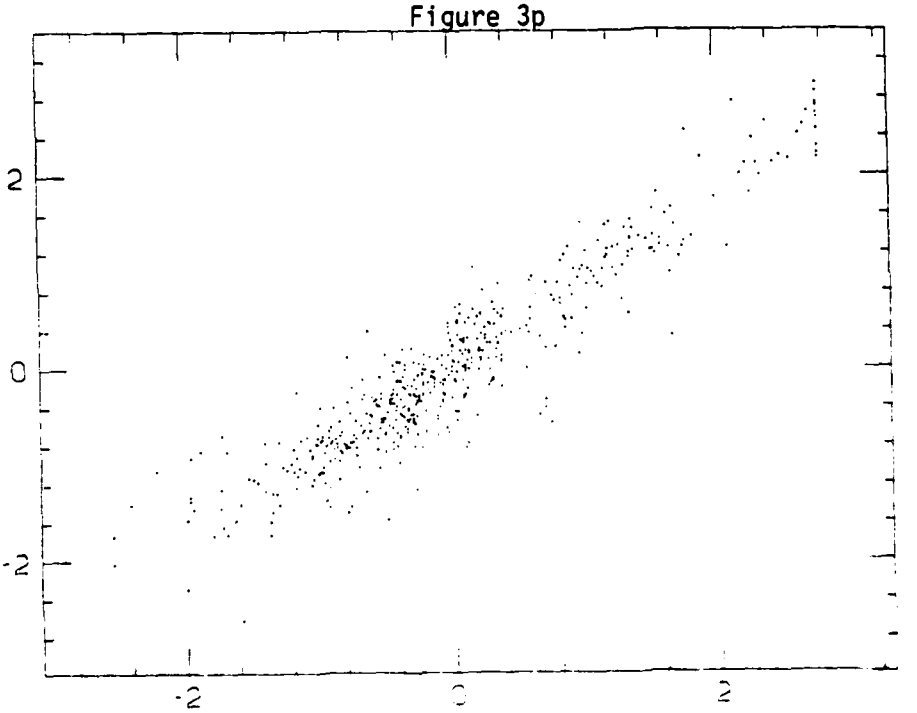


Figure 3p



4. Discussion

The ACE algorithm provides a fully automated method for estimating optimal transformations in multiple regression. It also provides a method for estimating maximal correlation between random variables. It differs from other empirical methods for finding transformations (Box and Tidwell [1962]; Anscombe and Tukey [1963]; Box and Cox [1964]; Kruskal [1965]; Draper and Cox [1969]; Fraser [1967]; Linsey [1972]; Box and Hill [1974]; Linsey [1974]; Wood [1974]; Mosteller and Tukey [1977]; and Tukey [1982]) in that the "best" transformations of the response and predictor variables are unambiguously defined and estimated without use of ad hoc heuristics, restrictive distributional assumptions, or restriction of the transformation to a particular parametric family.

The algorithm is reasonably computer efficient. On the Boston housing data set comprising 506 data points with 14 variables each, the run took 12 seconds of CPU time on an IBM 3081. Our guess is that this translates into 2.5 minutes on a VAX 11/750 with FP. To extrapolate to other problems, use the estimate that running time is proportional to (number of variables) $\times$ (sample size).

A strong advantage of the ACE procedure is the ability to incorporate variables of quite different type in terms of the set of values they can assume. The transformation functions $\theta(y),\phi_1(x_1),\ldots,\phi_p(x_p)$ assume values on the real line. Their arguments can, however, assume values on any set. For example, ordered real, periodic (circularly valued) real, ordered and unordered categorical variables can be incorporated in the same regression equation. For periodic variables, the smoother window need only wrap around the boundaries. For categorical variables, the procedure can be regarded as estimating optimal scores for each of their

values. (The special case of a categorical response and a single categorical predictor variable is known as canonical analysis--see Kendall and Stuart [1967], p.568--and the optimal scores can, in this case, also be obtained by solution of a matrix eigenvector problem.)

In some problems the analyst may wish to restrict $\hat{\theta}(y)$ to be monotone. For example, $\hat{\theta}(y)$ monotone allows the unique specification of y given a value of $\hat{\theta}(y)$. There is an option in the program (Appendix 2) that allows this using Kruskal's method of finding closest monotone fits (see Kruskal [1964], pp. 126-128). However, we advise running the regular algorithm first, since lack of monotonicity in $\hat{\theta}^*(y)$ can provide valuable insight into the structure of the data.

The solution functions $\hat{\theta}^*(y)$ and $\hat{\phi}_1^*(x_1),\ldots,\hat{\phi}_p^*(x_p)$ can be stored as a set of values associated with each observation $(y_k,x_{k1},\ldots,x_{kp})$, $1 \leq k \leq N$. However, since $\theta(y)$ and $\phi(x)$ are usually smooth (for continuous y, x), they can be easily approximated and stored as cubic spline functions (deBoor [1978]) with a few knots.

As a tool for data analysis, the ACE procedure provides graphical output to indicate a need for transformations, as well as to guide in their choice. If a particular plot suggests a familiar functional form for a transformation, it can be substituted for the empirical transformation estimate and the ACE algorithm can be rerun using an option which alters only the scale and origin of that particular transformation. The resulting e^2 can be compared to the original value. We have found that the plots themselves often give surprising new insights into the relationship between the response and predictor variables.

As with any regression procedure, a high degree of association between predictor variables can sometimes cause the individual transformation

estimates to be highly variable even though the complete model is reasonably stable. When this is suspected, running the algorithm on randomly selected subsets of the data, or on bootstrap samples (Efron [1979]) can assist in assessing the variability.

The ACE method has generality beyond that exploited here. An immediate generalization would involve multiple response variables $Y_1,\ldots,Y_q$. The generalized algorithm would estimate optimal transformations $\theta_1^*,\ldots,\theta_q^*,\phi_1^*,\ldots,\phi_p^*$ that minimize

$$E[\Sigma_{\ell=1}^{q}\theta_\ell(Y_\ell) - \Sigma_{j=1}^{p}\phi_j(X_j)]^2$$

subject to $E\theta_\ell = 0,\ \ell=1,\ldots,q,\ E\phi_j=0, j=1,\ldots,p$

and $\|\Sigma_{\ell=1}^{q}\theta_\ell(Y_\ell)\|^2 = 1$.

This extension generalizes the ACE procedure in a sense similar to that in which canonical correlation generalized linear regression.

The ACE algorithm (Section 2) is easily modified to incorporate this extension. An inner loop over the response variables, analagous to that for the predictor variables, replaces the single function minimization.

5.0 Optimal Transformations in Function Space

Introduction

Define random variables to take values either in the reals or in a finite or countable unordered set. Given a set of random variables $Y, X_1, \ldots, X_p$, a *transformation* is defined by a set of real valued measurable functions $(\theta, \phi_1, \ldots, \phi_p) = (\theta, \underline{\phi})$, each function defined on the range of the corresponding random variables, such that

$$(5.1) \qquad \begin{aligned} E\theta(Y) &= 0 , \quad E\phi_j(X_j) = 0 , \quad j = 1, \ldots, p \\ E\theta^2(Y) &< \infty , \quad E\phi_j^2(X_j) < \infty , \quad j = 1, \ldots, p \end{aligned}$$

Use the notation

$$(5.2) \qquad \tilde{\phi}(\underline{X}) = \sum_j^p \phi_j(X_j)$$

Denote the set of all transformations by F.

(5.3) DEFINITION. *A transformation* $(\theta^*, \underline{\phi}^*)$ *is optimal for regression if* $E(\theta^*)^2 = 1$, *and*

$$e^{*2} = E[\theta^*(Y) - \tilde{\phi}^*(\underline{X})]^2 = \inf_F \{E[\theta(Y) - \tilde{\phi}(\underline{X})]^2 ; E\theta^2 = 1\}$$

(5.4) DEFINITION. *A transformation* $(\theta^{**}, \underline{\phi}^{**})$ *is optimal for correlation if* $E(\theta^{**})^2 = 1$, $E(\tilde{\phi}^{**})^2 = 1$,

$$\rho^* = E[\theta^{**}(Y)\tilde{\phi}^{**}(\underline{X})] = \sup_F \{E[\theta(Y)\tilde{\phi}(\underline{X})]^2 ; E(\tilde{\phi})^2 = 1, E\theta^2 = 1\}$$

(5.5) THEOREM. *If* $(\theta^{**}, \underline{\phi}^{**})$ *is optimal for correlation, then* $\theta^* = \theta^{**}$, $\underline{\phi}^* = \rho^* \underline{\phi}^{**}$ *is optimal for regression and conversely. Furthermore* $e^{*2} = 1 - \rho^{*2}$.

PROOF. Write

$$E(\theta-\hat{\phi})^2 = 1 - E\theta\tilde{\phi} + E\tilde{\phi}^2$$
$$= 1 - 2E(\theta\tilde{\phi}^1)\sqrt{E\tilde{\phi}^2} + E\tilde{\phi}^2$$

where $\tilde{\phi} = \tilde{\phi}/\sqrt{E\tilde{\phi}^2}$. Hence

$$E(\theta-\tilde{\phi})^2 \geq 1 - 2\rho^*\sqrt{E\tilde{\phi}^2} + E\tilde{\phi}^2 \tag{5.6}$$

with equality only if $E\theta\tilde{\phi} = \rho^*$. The minimum of the right side of (5.6) over $E\tilde{\phi}^2$ is at $E\tilde{\phi}^2 = (\rho^*)^2$ where it is equal to $1-(\rho^*)^2$. Then $(e^*)^2 = 1-(\rho^*)^2$ and if $(\theta^{**},\underline{\phi}^{**})$ is optimal for correlation, then $\theta^* = \theta^{**}$, $\underline{\phi}^* = \rho^*\underline{\phi}^{**}$ is optimal for regression. The argument is reversible.

5.1 Existence of Optimal Transformations

To show existence of optimal transformations, two additional assumptions are needed:

AI. *There is no non-zero set of functions satisfying (5.1) such that*

$$\theta(Y) + \textstyle\sum_j \theta_j(X_j) = 0 \ a.s.$$

To formulate the second assumption, define

(5.7) DEFINITION. *Define the Hilbert spaces* $H_2(Y), H_2(X_1), \ldots, H_2(X_p)$ *as the sets of functions satisfying (5.1) with the usual inner product, i.e.,* $H_2(X_j)$ *is the set of all measurable* ϕ_j *such that* $E\phi_j(X_j) = 0$, $E\phi_j^2(X_j) < \infty$ *with* $(\phi_j', \phi_j) = E[\phi_j'(X_j)\phi_j(X_j)]$.

AII. *The conditional expectation operators*

$$E(\phi_j(X_j)|Y)\colon H_2(X_j) \to H_2(Y) ,$$
$$E(\phi_j(X_j)|X_i)\colon H_2(X_j) \to H_2(X_i) , \quad i \neq j$$
$$E(\theta(Y)|X_j)\colon H_2(Y) \to H_2(X_j)$$

are all compact.

Condition AII is satisfied in most cases of interest. A sufficient condition is given by: let X, Y be random variables with joint density $f_{X,Y}$ and marginals f_X, f_Y. Then the conditional expectation operator on $H_2(Y) \to H_2(X)$ is compact if

$$\iint [f_{XY}^2/f_X f_Y]dxdy < \infty \tag{5.8}$$

(5.9) THEOREM. *Under AI and AII optimal transformations exist.*

Some machinery is needed.

(5.10) PROPOSITION. *The set of all functions* f *of the form*

$$f(Y,\underline{X}) = \theta(Y) + \textstyle\sum_j \theta_j(X_j) , \quad \theta \in H_2(Y), \ \phi_j \in H_2(X_j)$$

with the inner product and norm

$$(g,f) = E[gf] , \quad \|f\|^2 = Ef^2 ,$$

is a Hilbert space denoted by H_2. *The subspace of all functions* $\tilde{\phi}$ *of the form*

$$\tilde{\phi}(\underline{X}) = \textstyle\sum_1^p \phi_j(X_j) , \quad \phi_j \in H_2(X_j)$$

is a closed linear subspace denoted by $H_2(X)$. *So are* $H_2(Y), H_2(X_1), \ldots, H_2(X_p)$.

(5.10) follows from:

(5.11) PROPOSITION. *Under AI, AII there are constants* $0 < c_1 \le c_2 < \infty$ *such that*

$$c_1(\|\theta\|^2 + \textstyle\sum_1^p \|\phi_j\|^2) \le \|\theta + \sum_1^p \phi_j\|^2 \le c_2(\|\theta\|^2 + \sum_1^p \|\phi_j\|^2) .$$

PROOF. The right hand inequality is immediate. If the left side does not hold, we can find a sequence $f_n = \theta_n + \sum \phi_{n,j}$ such that $\|\theta_n\|^2 + \sum_1^p \|\phi_{n'j}\|^2 = 1$, but $\|f_n\|^2 \to 0$. There is a subsequence n' such that $\theta_{n'} \xrightarrow{w} \theta$, $\phi_{n'j} \xrightarrow{w} \phi_j$ in the sense of weak convergence in $H_2(Y), H_2(X_1), \ldots, H_2(X_p)$ respectively.

Write

$$E[\phi_{n'j}(X_j)\phi_{n'i}(X_i)] = E[\phi_{n'j}(X_j)E(\phi_{n'i}(X_i)|X_j)]$$

to see that AII implies $E\phi_{n'j}\phi_{n'i} \to E\phi_j\phi_i$, $i \ne j$ and similarly for $E\theta_{n'}\phi_{n',j}$. Furthermore $\|\phi_i\| \le \underline{\lim}\, \|\phi_{n'i}\|$, $\|\theta\| \le \underline{\lim}\, \|\theta_{n'}\|$. Thus, defining $f = \theta + \sum_j \phi_j$

$$\|f\|^2 = \|\theta + \textstyle\sum_j \phi_j\| \le \underline{\lim}\, \|f_{n'}\|^2 = 0$$

which implies, by AI, that $\theta = \phi_1 = \cdots = \phi_p = 0$. On the other hand,

$$\|f_n\|^2 = \|\theta_{n'}\|^2 + \sum_j \|\phi_{n'j}\|^2 + \sum_j (\theta_{n'}, \phi_{n',j}) + \sum_{i \ne j} (\phi_{n'j}, \phi_{n',i})$$

Hence, if $f = 0$, then $\underline{\lim}\, \|f_n\|^2 \ge 1$.

(5.12) COROLLARY. *If* $f_n \xrightarrow{w} f$ *in* H_2, *then* $\theta_n \xrightarrow{w} \theta$ *in* $H_2(Y)$, $\phi_{nj} \xrightarrow{w} \phi_j$ *in* $H_2(X_j)$, $j = 1, \ldots, p$, *and conversely.*

PROOF. If $f_n = \theta_n + \sum_j \phi_{nj} \xrightarrow{w} \theta + \sum_j \phi_j$, then by (5.11), $\overline{\lim}\, \|\theta_n\| < \infty$, $\overline{\lim}\, \|\phi_{nj}\| < \infty$. Take n' such that $\theta_{n'} \xrightarrow{w} \theta'$, $\phi_{n'j} \xrightarrow{w} \phi_j'$, and let

$f' = \theta' + \sum_j \phi_j'$. Then for any $g \in H_2$, $(g, f_{n'}) \dashrightarrow (g', f')$, so $(g,f) = (g,f')$ all g. The converse is easier.

(5.13) DEFINITION. *In* H_2, *let* P_Y, P_j *and* P_X *denote the projection operators on* $H_2(Y)$, $H_2(X_j)$ *and* $H_2(X)$ *respectively.*

On $H_2(X_i)$, P_j, $j \neq i$, is the conditional expectation operator, and similarly for other subspaces.

(5.14) PROPOSITION. P_Y *is compact on* $H_2(X) \rightarrow H_2(Y)$ *and* P_X *is compact on* $H_2(Y) \rightarrow H_2(X)$.

PROOF. Take $\tilde{\phi}_n \in H_2(X)$, $\tilde{\phi}_n \xrightarrow{w} \tilde{\phi}$. This implies, by (5.12), that $\phi_{n,j} \xrightarrow{w} \phi_j$. By AII, $P_Y\phi_{n,j} \xrightarrow{s} P_Y\phi_j$ so that $P_Y\tilde{\phi}_n \xrightarrow{s} P_Y\tilde{\phi}$. Now take $\theta \in H_2(Y)$, $\tilde{\phi} \in H_2(X)$, then $(\theta, P_Y\tilde{\phi}) = (\theta, \tilde{\phi}) = (P_X\theta, \tilde{\phi})$. Thus, $P_X\colon H_2(Y) \rightarrow H_2(X)$ is the adjoint of P_Y and hence compact.

Now to complete the proof of Theorem 5.9. Consider the functional $\|\theta - \tilde{\phi}\|^2$ on the set of all $(\theta, \tilde{\phi})$ with $\|\theta\|^2 = 1$. For any θ, $\tilde{\phi}$

$$\|\theta - \tilde{\phi}\|^2 \geq \|\theta - P_X\theta\|^2$$

If there is a θ^* which achieves the minimum of $\|\theta - P_X\theta\|^2$ over $\|\theta\|^2 = 1$, then an optimal transformation is θ^*, $P_X\theta^*$. On $\|\theta\|^2 = 1$

$$\|\theta - P_X\theta\|^2 = 1 - \|P_X\theta\|^2 .$$

Let $\bar{s} = \{\sup \|P_X\theta\| ; \|\theta\| = 1\}$. Take θ_n such that $\|\theta_n\|^2 = 1$, $\theta_n \xrightarrow{w} \theta$, and $\|P_X\theta_n\| \rightarrow \bar{s}$. By the compactness of P_X, $\|P_X\theta_n\| \rightarrow \|P_X\theta\| = \bar{s}$. Further, $\|\theta\| \leq 1$. If $\|\theta\| < 1$, then for $\theta' = \theta/\|\theta\|$, we get the contradiction $\|P_X\theta'\| > \bar{s}$. Hence $\|\theta\| = 1$ and $(\theta, P_X\theta)$ is an optimal transformation.

5.2 Characterization of Optimal Transformations

Define two operators $U: H_2(Y) \to H_2(Y)$ and $V: H_2(X) \to H_2(X)$ by

$$U\theta = P_Y P_X \theta \ , \quad V\tilde{\phi} = P_X P_Y \tilde{\phi}$$

(5.15) PROPOSITION. *U and V are compact, self-adjoint and non-negative definite. They have the same eigenvalues and there is a 1-1 correspondence between eigenspaces for a given eigenvalue specified by*

$$\tilde{\phi} = P_X\theta / \|P_X\theta\| \ , \quad \theta = P_Y\tilde{\phi} / \|P_Y\tilde{\phi}\|$$

PROOF. Direct verification.

Let the largest eigenvalue be denoted by $\bar{\lambda}$, $\bar{\lambda} = \|U\| = \|V\|$. Then

(5.16) THEOREM. *If θ^*, $\tilde{\phi}^*$ is an optimal transformation for regression, then*

$$\bar{\lambda}\theta^* = U\theta^* \ , \quad \bar{\lambda}\tilde{\phi}^* = V\tilde{\phi}^*$$

Conversely, if θ satisfies $\bar{\lambda}\theta = U\theta$, $\|\theta\| = 1$, then θ, $P_X\theta$ is optimal for regression. If $\tilde{\phi}$ satisfies $\bar{\lambda}\tilde{\phi} = V\tilde{\phi}$, then $\theta = P_Y\tilde{\phi}/\|P_Y\tilde{\phi}\|$, and $\bar{\lambda}\tilde{\phi}/\|P_Y\tilde{\phi}\|$ are optimal for regression. In addition

$$(e^*)^2 = 1 - \bar{\lambda} \ .$$

PROOF. Let θ^*, $\tilde{\phi}^*$ be optimal. Then $\tilde{\phi}^* = P_X\theta^*$. Write

$$\|\theta^* - \tilde{\phi}^*\|^2 = 1 - 2(\theta^*, \tilde{\phi}^*) + \|\tilde{\phi}^*\|^2$$

Note that $(\theta^*, \tilde{\phi}^*) = (\theta^*, P_Y\tilde{\phi}^*) \leq \|P_Y\tilde{\phi}^*\|$ with equality only if $\theta^* = eP_Y\tilde{\phi}^*$, e constant. Therefore, $\theta^* = P_Y\tilde{\phi}^*/\|P_Y\tilde{\phi}^*\|$. This implies

$$\|P_Y\tilde{\phi}^*\|\theta^* = U\theta^* \ , \quad \|P_Y\tilde{\phi}^*\|\tilde{\phi}^* = V\tilde{\phi}^* \ ,$$

so that $\|P_Y\tilde{\phi}^*\|$ is an eigenvalue λ^* of U, V. Computing gives $\|\theta^*-\tilde{\phi}^*\|^2 = 1-\lambda^*$. Now take θ any eigenfunction of U corresponding to $\bar{\lambda}$, with $\|\theta\| = 1$. Let $\tilde{\phi} = P_X\theta$, then $\|\theta-\tilde{\phi}\| = 1-\bar{\lambda}$. This shows that θ^*, $\tilde{\phi}^*$ are not optimal unless $\lambda^* = \bar{\lambda}$. The rest of the theorem is straightforward verification.

(5.17) COROLLARY. *If $\bar{\lambda}$ has multiplicity one, then the optimal transformation is unique up to a sign change. In any case, the set of optimal transformations is finite dimensional.*

It appears that uniqueness is the general case.

5.3 Alternating Conditional Methods

Direct solution of the equations $\bar{\lambda}\theta = U\theta$ or $\bar{\lambda}\tilde{\phi} = V\tilde{\phi}$ is formidable. Attempting to use data to directly estimate the solutions is just as difficult. In the bivariate case, if X, Y are categorical, then $\bar{\lambda}\theta = U\theta$ becomes a matrix eigenvalue problem and is tractable. This is the case treated in Kendall and Stuart [1967].

The ACE algorithm is founded on the observation that there is an iterative method for finding optimal transformations. We illustrate this in the bivariate case. The goal is to minimize $\|\theta(Y)-\phi(X)\|^2$ with $\|\theta\|^2 = 1$. Denote $P_X\theta = E(\theta|X)$, $P_Y\phi = E(\phi|Y)$. Start with any first guess function $\theta_0(Y)$. Then define a sequence of functions by

$$\phi_0 = P_X\theta_0$$
$$\theta_1 = P_Y\phi_0/\|P_Y\phi_0\|$$
$$\phi_1 = P_X\theta_1$$

and in general $\phi_{n+1} = P_X\theta_n$, $\theta_{n+1} = P_Y\phi_{n+1}/\|P_Y\phi_{n+1}\|$. It is clear that at each step in the iteration $\|\theta-\phi\|^2$ is decreased. It is not hard to show that in general, θ_n, ϕ_n converge to an optimal transformation.

The above method of alternating conditionals extends to the general multivariate case. The analogue is clear; given θ_n, $\tilde{\phi}_n$, then the next iteration is

$$\tilde{\phi}_{n+1} = P_X\theta_n\ , \quad \theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|$$

However, there is an additional issue: How can $P_X\theta$ be computed using only the conditional expectation operators P_j, $j=1,\ldots,p$? This is done by starting with some function $\tilde{\phi}_0$ and iteratively subtracting off the projections of $\theta-\tilde{\phi}_n$ on the subspaces $H_2(X_1),\ldots,H_2(X_p)$ until we get a function $\tilde{\phi}$ such that the projection of $\theta-\tilde{\phi}$ on each of $H_2(X_j)$ is zero. This leads to

The Double Loop Algorithm

The Outer Loop

1. Start with an initial guess $\theta_0(Y)$.
2. Put $\tilde{\phi}_{n+1} = P_X\theta_n$, $\theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|$ and repeat until convergence.

Let $P_E\theta_0$ be the projection of θ_0 on the eigenspace E of U corresponding to $\bar{\lambda}$. Then

(5.18) THEOREM. *If* $\|P_E\theta_0\| \neq 0$, *define an optimal transformation by* $\theta^* = P_E\theta_0/\|P_E\theta_0\|$, $\tilde{\phi}^* = P_X\theta_0^*$. *Then* $\|\theta_n-\theta^*\| \to 0$, $\|\tilde{\phi}_n-\tilde{\phi}^*\| \to 0$.

PROOF. Notice that $\theta_{n+1} = U\theta_n/\|U\theta_n\|$. For any n, $\theta_n = \alpha_n\theta^* + g_n$, where $g_n \perp E$. Because, if it is true for n, then

$$\theta_{n+1} = (\alpha_n \bar{\lambda}\theta^* + Ug_n)/\|\alpha_n \bar{\lambda}\theta^* + Ug_n\|$$

and Ug_n is $\perp$ to E. For any $g \perp E$, $\|Ug\| \leq r\|g\|$ where $r < \bar{\lambda}$. Since $\alpha_{n+1} = \bar{\lambda}\alpha_n/\|U\theta_n\|$, $g_{n+1} = Ug_n/\|U\theta_n\|$, then

$$\|g_{n+1}\|/\alpha_{n+1} = \|Ug_n\|/\bar{\lambda}\alpha_n \leq (r/\bar{\lambda})\|g_n\|/\alpha_n .$$

Thus $\|g_n\|/\alpha_n \leq c(r/\lambda)^n$. But $\|\theta_n\| = 1$, $\alpha_n^2 + \|g_n\|^2 = 1$, implying $\alpha_n^2 \to 1$. Since $\alpha_0 > 0$, then $\alpha_n > 0$, so $\alpha_n \to 1$. Now use $\|\theta_n - \theta^*\| = (1-\alpha_n)^2 + \|g_n\|^2$ to reach the conclusion. Since $\|\tilde{\phi}_{n+1} - \tilde{\phi}^*\| = \|P_\chi\theta_n - P_\chi\theta^*\| \leq \|\theta_n - \theta^*\|$, the theorem follows.

The Inner Loop

1. Start with functions θ, $\tilde{\phi}_0$.
2. If, after m stages of iteration, the functions are $\phi_j^{(m)}$, then define, for $j = 1,2,\ldots,p$,

$$\phi_j^{(m+1)} = P_j(\theta - \sum_{i>j} \phi_j^{(m)} - \sum_{i<j} \phi_i^{(m+1)})$$

(5.19) THEOREM. *Let* $\tilde{\phi}_m = \sum_j \phi_j^{(m)}$. *Then* $\|P_\chi\theta - \tilde{\phi}_m\| \to 0$.

PROOF. Define the operator T by

$$T = (I-P_p)(I-P_{p-1})\cdots(I-P_1)$$

Then the iteration in the inner loop is expressed as

(5.20)
$$\begin{aligned} \theta - \phi_{m+1} &= T(\theta - \tilde{\phi}_m) \\ &= T^m(\theta - \tilde{\phi}_0) \end{aligned}$$

Write $\theta - \tilde{\phi}_0 = \theta - P_\chi\theta + P_\chi\theta - \tilde{\phi}_0$. Noting that $T(\theta - P_\chi\theta) = \theta - P_\chi\theta$, (5.20) becomes

$$\tilde{\phi}_{m+1} = P_\chi\theta + T^m(P_\chi\theta - \tilde{\phi}_0)$$

The theorem is then proven by

(5.21) PROPOSITION. *For any* $\tilde{\phi} \in H_2(X)$, $\|T^m\tilde{\phi}\| \to 0$.

PROOF. $\|(I-P_j)\tilde{\phi}\|^2 = \|\tilde{\phi}\|^2 - \|P_j\tilde{\phi}\|^2 \leq \|\tilde{\phi}\|^2$. Thus $\|T\| \leq 1$. There is no $\tilde{\phi} \neq 0$ such that $\|T\tilde{\phi}\| = \|\tilde{\phi}\|$. If there were, then $\|P_j\tilde{\phi}\| = 0$, all j. Then for $\tilde{\phi}' = \sum\phi_j'$,

$$(\tilde{\phi},\tilde{\phi}') = \sum_j(\tilde{\phi},\tilde{\phi}_j') = \sum_j(P_j\tilde{\phi},\phi_j') = 0$$

The operator $T = I + W$, where W is compact. Now we claim that $\|T^mW\| \to 0$ on $H_2(X)$. To prove this, let $0 < \gamma < 1$ and define

$$G(\gamma) = \sup_{\tilde{\phi}} \{\|TW\tilde{\phi}\|/\|W\tilde{\phi}\|;\ \|\tilde{\phi}\| \leq 1,\ \|W\tilde{\phi}\| \geq \gamma\} .$$

Take $\tilde{\phi}_n \xrightarrow{w} \tilde{\phi}$, $\|\tilde{\phi}_n\| \leq 1$, $\|W\tilde{\phi}_n\| \geq \gamma$ so that $\|TW\tilde{\phi}_n\|/\|W\tilde{\phi}_n\| \to G(\gamma)$. Then $\|\tilde{\phi}\| \leq 1$, $\|W\tilde{\phi}\| \geq \gamma$ and $G(\gamma) = \|TW\tilde{\phi}\|/\|W\tilde{\phi}\|$. Thus $G(\gamma) < 1$, for all $0 < \gamma < 1$ and is clearly non-increasing in γ. Then

$$\|T^mW\tilde{\phi}\| = \|TWT^{m-1}\tilde{\phi}\| \leq G(\|T^{m-1}W\tilde{\phi}\|)\|T^{m-1}W\tilde{\phi}\| .$$

Put $\gamma_0 = \|W\|$, $\gamma_m = G(\gamma_{m-1})\gamma_{n-1}$, then $\|T^mW\| \leq \gamma_m$. But clearly $\gamma_m \to 0$.

The range of W is dense in $H_2(X)$. Otherwise, there is a $\tilde{\phi}' \neq 0$ such that $(\tilde{\phi}',W\tilde{\phi}) = 0$, all $\tilde{\phi}$. This implies $(W^*\tilde{\phi}',\tilde{\phi}) = 0$ or $W^*\tilde{\phi}' = 0$. Then $\|T^*\tilde{\phi}'\| = \|\tilde{\phi}'\|$ and a repetition of the argument given above leads to $\tilde{\phi}' = 0$. For any $\tilde{\phi}$ and $\varepsilon > 0$, take $W\tilde{\phi}_1$ so that $\|\tilde{\phi}-W\tilde{\phi}_1\| \leq \varepsilon$. Then $\|T^m\tilde{\phi}\| \leq \varepsilon + \|T^mW\tilde{\phi}_1\|$, which completes the proof.

There are two versions of the double loop. In the first, the initial functions $\tilde{\phi}_0$ are the limiting functions produced by the preceding inner loop. This is called the *restart version*. In the second, the

initial functions are $\tilde{\phi}_0 \equiv 0$. This is the *fresh start* version. The main theoretical difference is that a stronger consistency result holds for fresh start. Restart is a faster running algorithm, and is embodied in the ACE code.

The Single Loop Algorithm

The single loop algorithm combines a single iteration of the inner loop with an iteration of the outer loop. Thus, it is summarized by

1. Start with θ_0, $\tilde{\phi}_0 = 0$.
2. If the current functions are θ_n, $\tilde{\phi}_n$, define $\tilde{\phi}_{n+1}$ by

$$\theta_n - \tilde{\phi}_{n+1} = T(\theta_n - \tilde{\phi}_n)$$

3. Let $\theta_{n+1} = P_Y\tilde{\phi}_{n+1}/\|P_Y\tilde{\phi}_{n+1}\|$. Run to convergence.

This is a cleaner algorithm than the double loop and its implementation on data runs at least twice as fast as the double loop and requires only a single convergence test. Unfortunately, we have been unable to prove that it converges in function space. Assuming convergence, it can be shown that the limiting θ is an eigenfunction of U. But giving conditions for θ to correspond to $\bar{\lambda}$ or even showing that θ will correspond to $\bar{\lambda}$ "almost always" seems difficult.

6.0 The ACE Algorithm on Finite Data Sets

Introduction

The ACE algorithm is implemented on finite data sets by replacing conditional expectations, given continuous variables, by data smooths. In looking at the convergence and consistency properties of the ACE algorithm, the critical element was the properties of the data smooth used. The results are fragmentary. Convergence of the algorithm is proven only for a very restricted class of smooths. In practice, in over 1000 runs of ACE over a wide variety of data sets and using three different types of smooths, we have seen only one instance of failure to converge. A fairly general, but weak, consistency proof is given. We conjecture the form of a stronger consistency result.

6.1 Data Smooths

Define a data set D to be a set $\{\underline{x}_1,\ldots,\underline{x}_N\}$ of N points in p dimensional space, i.e. $\underline{x}_k = (x_{k1},\ldots,x_{kp})$. Let $\mathcal{D}_N$ be the collection of all such data sets. For fixed D, define $F(\underline{x})$ as the space of all real-valued functions ϕ defined on D, i.e. $\phi \in F(\underline{x})$ is defined by the N real numbers $\{\phi(\underline{x}_1),\ldots,\phi(\underline{x}_N)\}$. Define $F(x_j)$, $j = 1,\ldots,p$ as the space of all real-valued functions defined on the set $\{x_{1j},x_{2j},\ldots,x_{nj}\}$.

(6.1) DEFINITION. *A data smooth* S *of* $\underline{x}$ *on* x_j *is a mapping* $S\colon F(\underline{x}) \rightarrow F(x_j)$ *defined for every* D *in* $\mathcal{D}_n$. *If* $\phi \in F(\underline{x})$ *denote the corresponding element in* $F(x_j)$ *by* $S(\phi|x_j)$ *and its values by* $S(\phi|x_{kj})$.

Let x be any one of $x_1,\ldots,x_p$. Some examples of data smooths are

1. Histogram: Divide the real axis up into disjoint intervals $\{I_\ell\}$. If $x_k \in I_\ell$, define

$$S(\phi|x_k) = \frac{1}{n_\ell} \sum_{\underline{x}_m;\ x \in I_\ell} \phi(\underline{x}_m)$$

2. Moving Average: Fix $M < N/2$. Order the x_k getting $x_1 < x_2 < \cdots < x_N$ (assume no ties), and corresponding $\phi(\underline{x}_1),\ldots,\phi(\underline{x}_n)$. Put

$$S(\phi|x_k) = \frac{1}{2M} \sum_{m \neq 0}^{M}{}_{-M} \phi(\underline{x}_{k+m})$$

If M points are not available on one side, make up the deficiency on the other side.

3. Kernel: Take $K(x)$ defined on the reals with maximum at $x = 0$. Then

$$S(\phi|x_k) = \sum_m \phi(\underline{x}_m)K(x_m - x_k) \Big/ \sum_\ell K(x_\ell - x_k)$$

4. Regression: Fix M and order x_k as in (2) above. At x_k, regress the values of $\phi(\underline{x}_{k-M}),\ldots,\phi(\underline{x}_{k+M})$ excluding $\phi(\underline{x}_k)$ on $x_{k-M},\ldots,x_{k+M}$ excluding x_k, getting a regression line $L(x)$. Put $S(\phi|x_k) = L(x_k)$. If M points are not available on each side of x_k make up the deficiency on the other side.

5. Supersmoother: See Friedman and Stuetzle [1982].

Some properties that are relevant to the behavior of smoothers are given below. These properties hold only if they are true for all $D \in \mathcal{D}_n$.

Linearity. A smooth is linear if

$$S(\alpha\phi_1 + \beta\phi_2) = \alpha S\phi_1 + \beta S\phi_2$$

for all ϕ_1, $\phi_2 \in F(\underline{x})$ and all constants α, β.

Constant Preserving. If $\phi \in F(\underline{x})$ is constant, $\phi \equiv c$, then $S\phi \equiv c$.

To give a further property, introduce the inner product $(\ )_N$ on $F(\underline{x})$ defined by

$$(\phi,\phi')_N = \frac{1}{n} \sum_k \phi(\underline{x}_k)\phi'(\underline{x}_k)$$

and the corresponding norm $\| \|_N$.

Boundedness. S is bounded by M if

$$\|S\phi\|_N \leq M\|\phi\|_N , \quad \text{all } \phi \in F(\underline{x})$$

In the examples of smooths given above, all are linear, except supersmoother. This implies they can be represented as an $N \times N$ matrix operator varying with D. All are constant preserving. Histograms and moving average are bounded by one. Regression is unbounded due to end effects, but in the appendix we introduce a modified regression smooth that is bounded by 2. Supersmoother is bounded by 2. The bound for kernel smooths is more complicated.

6.2 Convergence of ACE

Let the data be of the form $(y_k,\underline{x}_k) = (y_k,x_{k1},\ldots,x_{kp})$, $k = 1,\ldots,N$. Assume that $\bar{y} = \bar{x}_1 = \cdots = \bar{x}_p = 0$. Define smooths $S_y,S_1,\ldots,S_p$ where S_y: $F(y,\underline{x}) \rightarrow F(y)$ and S_j: $F(y,\underline{x}) \rightarrow F(x_j)$. Let $H_2(y,\underline{x})$ be the set of all functions in $F(y,\underline{x})$ with zero mean and $H_2(y)$, $H_2(x_j)$ the corresponding subspaces.

It is essential to modify the smooths so that the resulting functions have zero means. This is done by subtracting the mean; thus the modified

S_j is defined by

$$S_j\phi = S_j\phi - \overline{S_j\phi} \tag{6.2}$$

Henceforth, we *use only modified smooths* and assume the original smooth to be constant preserving, so that the modified smooths take constants into zero.

The ACE algorithm is defined by

1. $\theta^{(0)}(y_k) = y_k, \quad \phi_j^{(0)}(x_{kj}) \equiv 0.$

The Inner Loop

2. At the n stage of the outer loop, start with $\theta^{(n)}, \phi_j^{(0)}$.
   For every $m \geq 1$ and $j = 1,\ldots,p$ define

$$\phi_j^{(m+1)} = S_j(\theta^{(n)} - \sum_{i<j} \phi_i^{(m+1)} - \sum_{i>j} \phi_i^{(m)})$$

   Keep increasing m until convergence to ϕ_j.

The Outer Loop

3. Set $\theta^{(n+1)} = S_y(\sum_j \phi_j)/\|S_y(\sum_j \phi_j)\|_N$, go back to the inner loop with $\phi_j^{(0)} = \phi_j$. Continue until convergence.

To formalize this algorithm, introduce the space $H_2(\theta,\underline{\phi})$ with elements $(\theta,\phi_1,\ldots,\phi_p)$, $\theta \in H_2(y)$, $\phi_j \in H_2(x_j)$, and subspaces $H_2(\theta)$ with elements $(\theta,0,0,\ldots,0) = \underline{\theta}$ and $H_2(\underline{\phi})$ with elements $(0,\phi_1,\ldots,\phi_p) = \underline{\phi}$.

For $f = (f_0,f_1,\ldots,f_p)$ in $H_2(\theta,\underline{\phi})$ define $S_j\colon H_2(\theta,\underline{\phi}) \to H_2(\theta,\underline{\phi})$ by

$$(S_j f)_i = \begin{cases} 0, & j \neq i \\ f_j + S_j(\sum_{i\neq j} f_i), & j = i \end{cases}$$

Starting with $\underline{\theta} = (\theta,0,0,\ldots,0)$, $\underline{\phi}^{(m)} = (0,\phi_1^{(m)},\ldots,\phi_p^{(m)})$ one complete cycle in the inner loop is described by

$$\underline{\theta} - \underline{\phi}^{(m+1)} = (I-S_p)(I-S_{p-1})\cdots(I-S_1)(\underline{\theta}-\underline{\phi}^{(m)}) \quad . \tag{6.3}$$

Define $\hat{T}$ on $H_2(\theta,\underline{\phi}) \rightarrow H_2(\theta,\underline{\phi})$ as the product operator in (6.3). Then

$$\underline{\phi}^{(m)} = \underline{\theta} - \hat{T}^m(\underline{\theta}-\underline{\phi}^{(0)}) \quad . \tag{6.4}$$

If, for a given $\underline{\theta}$, the inner loop converges, then the limiting $\underline{\phi}$ satisfies

$$S_j(\underline{\theta}-\underline{\phi}) = 0 \quad , \quad j = 1,\ldots,p \quad . \tag{6.5a}$$

That is, the smooth of the residuals on any predictor variable is zero.

Adding

$$\underline{\theta} = S_y\underline{\phi}/\|S_y\underline{\phi}\|_N \tag{6.5b}$$

to (6.5a) gives a set of equations satisfied by the estimated optimal transformations.

Assume, for the remainder of this section, that the smooths are linear. Then (6.5a) can be written as

$$S_j\underline{\phi} = S_j\underline{\theta} \quad , \quad j = 1,\ldots,p \quad . \tag{6.6}$$

Let $sp(S_j)$ denote the spectrum of the matrix S_j. Assume $1 \notin sp(S_j)$. (For constant preserving smooths 1 is always in the spectrum, but not for modified smooths.) Define matrices A_j by $A_j = S_j(I-S_j)^{-1}$ and the matrix A as $\sum_j A_j$. Assume further that $-1 \notin sp(A)$. Then (6.6) has the unique solution

$$\phi_j = A_j(I+A)^{-1}\theta \quad , \quad j = 1,\ldots,p \tag{6.7}$$

The element $\underline{\phi} = (0,\phi_1,\ldots,\phi_p)$ given by (6.7) will be denoted by $\hat{P}\underline{\theta}$. Rewrite (6.3) using $(I-\hat{T})(\underline{\theta}-\hat{P}\underline{\theta}) = 0$ as

$$\underline{\phi}^{(m)} = \hat{P}\theta - \hat{T}^m(\hat{P}\theta-\underline{\phi}^{(0)}) \tag{6.8}$$

Therefore, the inner loop converges if it can be shown that $\hat{T}^m f \to 0$ for all $f \in H_2(\underline{\phi})$. What we can show is

(6.9) THEOREM. *If* $\det[I+A] \neq 0$ *and if the spectral radii of* $S_1,\ldots,S_p$ *are all less than one, a necessary and sufficient condition for* $\hat{T}^m f \to 0$ *for all* $f \in H_2(\underline{\phi})$ *is that*

$$\det[\lambda I - \prod_1^p (I-S_j/\lambda)^{-1}(I-S_j)] \tag{6.10}$$

have no zeroes in $|\lambda| \geq 1$ *except* $\lambda = 1$.

PROOF. For $\hat{T}^m f \to 0$, all $f \in H_2(\underline{\phi})$, it is necessary and sufficient that the spectral radius of $\hat{T}$ be less than one. The equation $\hat{T}f = \lambda f$ in component form is

$$\lambda f_j = -S_j(\lambda \sum_{i<j} f_i + \sum_{i>j} f_i) , \quad j = 1,\ldots,p . \tag{6.11}$$

Let $s = \sum_i f_i$ and rewrite (6.11) as

$$(\lambda I-S_j)f_j = S_j((1-\lambda)\sum_{i<j} f_i - s) . \tag{6.12}$$

If $\lambda = 1$, (6.12) becomes $(I-S_j)f_j = -S_j s$ or $s = -As$. By assumption, this implies $s = 0$, and hence $f_j = 0$, all j. This rules out $\lambda = 1$ as an eigenvalue of $\hat{T}$. For $\lambda \neq 1$, but λ greater than the maximum of the spectral radii of the S_j, $j = 1,\ldots,p$, define $g_j = (1-\lambda)\sum_{i<j} f_i - s$. Then $f_j = (g_{j+1}-g_j)/(1-\lambda)$, so

$$(\lambda I-S_j)(g_{j+1}-g_j) = (1-\lambda)S_j g_j$$

or

$$g_{j+1} = (I-S_j/\lambda)^{-1}(I-S_j)g_j \ . \tag{6.13}$$

Since $g_{p+1} = -\lambda s$, $g_1 = -s$, then (6.13) leads to

$$\lambda s = (I-S_p/\lambda)^{-1}(I-S_p)\cdots(I-S_1/\lambda)^{-1}(I-S_1)s \tag{6.14}$$

If (6.14) has no non-zero solutions, then $s = 0$, $g_j = 0$, $j = 1,\ldots,p$, implying all $f_j = 0$. Conversely, if (6.14) has a solution $s \neq 0$, it leads to a solution of (6.11).

Unfortunately, condition (6.10) is difficult to verify for general linear smooths. If the S_j are self-adjoint, non-negative definite, such that all elements in the unmodified smooth matrix are non-negative, then all spectral radii of S_j are less than one, and (6.10) can be shown to hold by verifying that

$$|\lambda| \leq \prod_1^p \|(I-S_j/\lambda)^{-1}(I-S_j)\|$$

has no solutions λ with $|\lambda| > 1$, and then ruling out solutions with $|\lambda| = 1$.

The only common type of smooth satisfying the above conditions is the histogram smooth, a poor smooth to use in implementing ACE.

Assuming that the inner loop converges to $\hat{P}\theta$, then the outer loop iteration is given by

$$\underline{\theta}^{(n+1)} = \frac{S_y \hat{P}\underline{\theta}^{(n)}}{\|S_y \hat{P}\underline{\theta}^{(n)}\|_N}$$

Put the matrix $S_y\hat{P} = \hat{U}$, so that

$$\theta^{(n+1)} = \frac{\hat{U}\theta^{(n)}}{\|\hat{U}\theta^{(n)}\|_N} \tag{6.15}$$

If the eigenvalue $\hat{\lambda}$ of $\hat{U}$ having largest absolute value is real and positive, then $\theta^{(n+1)}$ converges to the projection of $\theta^{(0)}$ on the eigenspace of $\hat{\lambda}$. The limiting θ, $\hat{P}\theta$ is a solution of (6.5a,b). However, if $\hat{\lambda}$ is not real and positive, then $\theta^{(n)}$ oscillates and does not converge. If the smooths are self-adjoint and non-negative definite, then $S_y\hat{P}$ is the product of two self-adjoint non-negative definite matrices, hence has only real non-negative eigenvalues. We are unable to find conditions guaranteeing this for more general smooths.

Thus, in spite of the fact that ACE has invariably converged (with one exception) we cannot give a general convergence proof. We conjecture that convergence holds only for "most" data sets in a sense made explicit in the following section.

6.3 Consistency of ACE

For $\phi_0,\phi_1,\ldots,\phi_p$ any functions in $H_2(Y),H_2(X_1),\ldots,H_2(X_p)$, and any data set $D \in \mathcal{D}_N$, define functions $P_j(\phi_i|x_j)$ by

$$P_j(\phi_i|x_{kj}) = E(\phi_i(X_i)|X_j = x_{kj}) \tag{6.16}$$

Let ϕ_i in $H_2(x_j)$ be defined as the restriction of ϕ_i to the set of data values $\{x_{1j},\ldots,x_{Nj}\}$ minus its mean value over the data values.

Assume that the N data vectors $(y_k,\underline{x}_k)$ are independent samples from the distribution of $(Y,X_1,\ldots,X_p)$.

(6.17) DEFINITION. *Let* $S_y^{(N)}$, $S_j^{(N)}$ *be any sequence of data smooths. They are mean square consistent if*

$$E\|S_j^{(N)}(\sum_{i\neq j}\phi_i|x_j) - \sum_{i\neq j}P_j(\phi_i|x_j)\|_N^2 \to 0$$

for all $\phi_0,\ldots,\phi_p$ *as above, with the analogous definition for* $S_y^{(N)}$.

The m.s. consistency of some smooths is discussed in the next section.

Whether or not the algorithm converges, a weak consistency result can be given under general conditions. Start with $\theta_0 \in H_2(Y)$. On each data set, run the inner loop iteration m times, that is, define

$$\underline{\phi}_m^{(n+1)} = \theta^{(n)} - \hat{T}^m(\theta^{(n)} - \underline{\phi}_m^{(n)})$$

Then set

$$\underline{\theta}_m^{(n+1)} = P_y\,\underline{\phi}_m^{(n+1)}/\|P_y\phi_m^{(n+1)}\|_N$$

Repeat the outer loop ℓ times getting the final functions $\bar{\theta}_N(y;m,\ell)$, $\phi_{jN}(x_j;m,\ell)$. Do the analogous thing in function space starting with θ_0, getting functions whose restriction to the data set D are denoted by $\theta(y;m,\ell)$, $\phi_j(x_j;m,\ell)$. Then

(6.18) THEOREM. *If the smooths* $S_y^{(N)}$, $S_j^{(N)}$ *are m.s. consistent and uniformly bounded as* $N\to\infty$, *then*

$$E\|\theta_N(y;m,\ell)-\theta(y;,m,\ell)\|_N^2 \to 0\ , \quad E\|\phi_{jN}(x_j;m,\ell)-\phi_j(x_j;m,\ell)\|_N^2 \to 0$$

If θ^* *is the optimal transformation* $P_E\theta_0/\|P_E\theta_0\|$, $\tilde{\phi}^* = P_X\theta^*$, *then as* $m,\ \ell\to\infty$ *in any way, for the fresh start algorithm*

$$\|\theta(\cdot;m,\ell)-\theta^*\| \to 0\ , \quad \|\phi_j(\cdot;m,\ell)-\phi_j^*\| \to 0\ .$$

PROOF. First note that for any product of smooths $S_{i_1}^{(N)} \cdots S_{i_\ell}^{(N)}$,

$$E\left\| S_{i_1}^{(N)} \cdots S_{i_\ell}^{(N)} \varepsilon_0 - P_{i_1} \cdots P_{i_\ell} \theta_0 \right\|_N^2 \to 0 .$$

This is illustrated with $S_i^{(N)} S_j^{(N)} \theta_0$, $i \neq j$. Since $E\|S_j^{(N)}\theta_0 - P_j\theta_0\|_N^2 \to 0$, then $S_j^{(N)}\theta_0 = P_j\theta_0 + \phi_{j,N}$ where $E\|\phi_{j,N}\|_N^2 \to 0$. Therefore

$$S_i^{(N)}(S_j^{(N)}\theta_0) = S_i^{(N)} P_j \theta_0 + S_i^{(N)} \phi_{j,N}$$

By assumption $\|S_i^{(N)}\phi_{j,N}\|_N \leq M\|\phi_{j,N}\|_N$, where M does not depend on N. Therefore $E\|S_i^{(N)}\phi_{j,N}\|_N^2 \to 0$. By assumption $E\|S_i^{(N)} P_j\theta_0 - P_i P_j \theta_0\|_N^2 \to 0$ so that $E\|S_i^{(N)} S_j^{(N)}\theta_0 - P_i P_j \theta_0\|_N^2 \to 0$.

(6.19) PROPOSITION. *If* θ_N *is defined in* $H_2(y)$ *for all data sets* D, *and* $\theta \in H_2(Y)$ *such that*

$$E\|\theta_N(y) - \theta(y)\|^2 \to 0$$

then

$$E\left\| \frac{\theta_N(y)}{\|\theta_N\|_N} - \frac{\theta(y)}{\|\theta\|} \right\|_N^2 \to 0 .$$

PROOF. Write $\theta/\|\theta\| = \theta/\|\theta\|_N + \theta(1/\|\theta\| - 1/\|\theta\|_N)$. So two parts are needed. First, to show that

$$E\left\| \frac{\theta_N}{\|\theta_N\|_N} - \frac{\theta}{\|\theta\|_N} \right\|_N^2 \to 0 .$$

Second, that $E\|\theta(\frac{1}{\|\theta\|} - \frac{1}{\|\theta\|_N})\|_N^2 \to 0$. For the first part, let

$$S_N^2 = \frac{1}{N} \sum_k \left(\frac{\theta_N(y_k)}{\|\theta_N\|_N} - \frac{\theta(y_k)}{\|\theta\|_N} \right)^2 = 2\left(1 - \frac{(\theta_N, \theta)_N}{\|\theta_N\|_N \|\theta\|_N}\right) .$$

Then $S_N^2 \leq 4$, so it is enough to show that $S_N^2 \xrightarrow{P} 0$ to get $ES_N^2 \to 0$.

Let

$$V_N^2 = \frac{1}{N}\sum_k (\theta_N(y_k)-\theta(y_k))^2$$

$$= \|\theta_N\|_N^2 + \|\theta\|_N^2 - 2(\theta_N,\theta)_N$$

$$= (\|\theta_N\|-\|\theta\|_N)^2 + 2(\|\theta\|_N\|\theta_N\|_N-(\theta_N,\theta)_N)$$

Both terms are positive, and since $EV_N^2 \to 0$, $E(\|\theta_N\|-\|\theta\|_N)^2 \to 0$, $E(\|\theta\|_N\|\theta_N\|_N-(\theta_N,\theta)_N) \to 0$. By the law of large numbers $E\left|\|\theta\|_N^2-\|\theta\|^2\right| \to 0$, resulting in $S_N^2 \xrightarrow{P} 0$.

Now look at

$$W_N^2 = \frac{1}{N}\sum_k \theta^2(y_k)\left[\frac{1}{\|\theta\|_N}-\frac{1}{\|\theta\|}\right]^2$$

$$= \|\theta\|_N^2\left(\frac{1}{\|\theta\|_N}-\frac{1}{\|\theta\|}\right)^2$$

$$= \left(1-\frac{\|\theta\|_N}{\|\theta\|}\right)^2$$

Then $EW_N^2 \to 0$ follows from $E\left|\|\theta\|_N^2-\|\theta\|^2\right| \to 0$.

Using Proposition 6.19 it follows that $E\|\theta_N(y;m,\ell)-\theta(y;m,\ell)\|_N^2 \to 0$ and in consequence, that $E\|\phi_{j,N}(x_j;m,\ell)-\phi_j(x_j;m,\ell)\|^2 \to 0$.

In function space, define

$$P_X^{(m)} = \theta - T^m\theta$$

$$U_m = P_Y P_X^{(m)}$$

Then

$$\theta(\cdot;m,\ell) = \frac{U_m^\ell\theta_0}{\|U_m^\ell\theta_0\|} .$$

The last step in the proof is showing that

$$\left\|\frac{U_m^\ell\theta_0}{\|U_m^\ell\theta_0\|} - \theta^*\right\| \to 0$$

as m, ℓ go to infinity. Begin with

(6.20) PROPOSITION. *As* $m \to \infty$, $U_m \to U$ *in the uniform operator norm.*

PROOF. $\|U_m\theta - U\theta\| = \|P_Y T^m P_X \theta\| \leq \|T^m P_X \theta\|$. Now on $H_2(Y)$, $\|T^m P_X\| \to 0$. If not, take θ_m, $\|\theta_m\| = 1$ such that $\|T^m P_X \theta_m\| \geq \delta$, all m. Let $\theta_{m'} \xrightarrow{w} \theta$, then $P_X\theta_m \xrightarrow{s} P_X\theta$, and

$$\begin{aligned}\|T^{m'} P_X \theta_{m'}\| &\leq \|T^{m'} P_X(\theta_{m'} - \theta)\| + \|T^{m'} P_X \theta\| \\ &\leq \|P_X(\theta_{m'} - \theta)\| + \|T^{m'} P_X \theta\|\end{aligned}$$

By Proposition (5.21) the right hand side goes to zero.

The operator U_m is not necessarily self-adjoint, but it is compact. By (6.18), if $O(sp(U))$ is any open set containing $sp(U)$, then for m sufficiently large $sp(U_m) \subset O(sp(U))$. Suppose, for simplicity, that the projection E_λ corresponding to the largest eigenvalue λ of U is one-dimensional. (The proof goes through if E_λ is higher-dimensional but is more complicated.) Then for any open neighborhood O of λ, and m sufficiently large, there is only one eigenvalue λ_m of U_m in O, $\lambda_m \to \lambda$ and the projection $P_E^{(m)}$ of U_m corresponding to λ_m converges to P_{E_λ} in the uniform operator topology. Also, λ_m can be taken as the eigenvalue of U_m having largest absolute value. If λ' is the second largest eigenvalue of U, and λ_m' the eigenvalue of U_m having the second highest absolute value, then (assuming $E_{\lambda'}$ is one-dimensional) $\lambda_m' \to \lambda'$.

Write

$$W_m = U_m - P_E^{(m)}, \quad W = U - P_{E_\lambda}$$

so again $\|W_m - W\| \to 0$. Now

(6.21)
$$U_m^{\ell}\theta_0 = \lambda_m^{\ell} P_E^{(m)}\theta_0 + W_m^{\ell}\theta_0$$
$$W^{\ell}\theta_0 = \lambda^{\ell} P_{E_\lambda}\theta_0 + W^{\ell}\theta_0$$

For any $\varepsilon > 0$ we will show that there exists m_0, ℓ_0 such that for $m \geq m_0$, $\ell \geq \ell_0$

(6.22)
$$\|W_m^{\ell}\theta_0\|/\lambda_m^{\ell} \leq \varepsilon \ , \quad \|W^{\ell}\theta_0\|/\lambda^{\ell} \leq \varepsilon \ .$$

Take $r = (\lambda+\lambda')/2$ and select m_0 such that $r > \max(\lambda', |\lambda_m'|, m \geq m_0)$. Denote by $R(\lambda, W_m)$ the resolvent of W_m. Then

$$W_m^{\ell} = \frac{1}{2\pi i}\int_{|\lambda|=r} \lambda^{\ell} R(\lambda, W_m)\, d\lambda$$

and

$$\|W_m^{\ell}\| \leq \frac{1}{2\pi} r^{\ell} \int_{|\lambda|=r} \|R(\lambda, W_m)\|\, d|\lambda|$$

where $d|\lambda|$ is arc length along $|\lambda| = r$. On $|\lambda| = r$, for $m \geq m_0$, $\|R(\lambda, W_m)\|$ is continuous and bounded. Furthermore $\|R(\lambda, W_m)\| \longrightarrow \|R(\lambda, W)\|$ uniformly. Letting $M(r) = \max_{|\lambda|=r} \|R(\lambda, W)\|$, then

$$\|W_m^{\ell}\| \leq r^{\ell} M(r)(1+\delta_m \Delta_m)$$

where $\Delta_m \delta_m \longrightarrow 0$ as $m \longrightarrow \infty$. Certainly

$$\|W^{\ell}\| \leq r^{\ell} M(r) \ .$$

Fix $\delta > 0$ such that $(1+\delta)r < \lambda$. Take m_0' such that for $m \geq \max(m_0, m_0')$, $\lambda_m \geq (1+\delta)r$. Then

$$\|W_m^{\ell}\|/\lambda_m^{\ell} \leq \left(\frac{1}{1+\delta}\right)^{\ell} M(r)(1+\Delta_m)$$

and

$$\|W^{\ell}\|/\lambda^{\ell} \leq \left(\frac{1}{1+\delta}\right)^{\ell} M(r) \ .$$

Now choose a new m_0 and ℓ_0 such that (5.6) is satisfied.

Using (5.6)

$$\left\| \frac{U_m^{\ell}\theta_0}{\|U_m^{\ell}\theta_0\|} - \frac{P_{E_m}\theta_0}{\|P_{E_m}\theta_0\|} \right\| = \varepsilon_{m,\ell}$$

where $\varepsilon_{m,\ell} \to 0$ as $m,\ell \to \infty$. Thus

$$\left\| \frac{U_m^{\ell}\theta_0}{\|U_m^{\ell}\theta_0\|} - \theta^* \right\| = \varepsilon'_{m,\ell} + \left\| \frac{P_{E_m}\theta_0}{\|P_{E_m}\theta_0\|} - \frac{P_{E_\lambda}\theta_0}{\|P_{E_\lambda}\theta_0\|} \right\|$$

and the right side goes to zero as $m,\ell \to \infty$.

The term weak consistency is used above because we have in mind a desirable stronger result. We conjecture that for reasonable smooths, the set $C_N = \{(Y_1,\underline{X}_1),\ldots,(Y_N,\underline{X}_N);$ algorithm converges$\}$ satisfies $P(C_N) \to 1$ and that for θ_N the limit on C_N starting from a fixed θ_0,

$$E[I_{C_N}\|\theta_N-\theta^*\|_N^2] \to 0 \ .$$

We also conjecture that such a theorem will be difficult to prove.

APPENDIX 1

Most Reasonable Sequences of Uniformly Bounded Smooths are M.S. Consistent

If the window size goes to zero at the right rate as $N \to \infty$, then most "reasonable" smooths which utilize local smoothing are m.s. consistent. There is a substantial literature on consistency, usually in higher dimensional spaces. Stone's pioneering paper [1977] established consistency for k-nearest neighbor smoothing. Devroye and Wagner [1980] and independently Spiegelman and Sacks [1980] gave weak conditions for consistency of kernel smooths. See Stone [1977] and Devroye [1981] for a review of the literature.

The common definition of consistency is: given a set of $N-1$ independent copies $(X_1,Y_1),\ldots,(X_{N-1},Y_{N-1})$, of (X,Y) drawn from the same bivariate distribution, and $\phi \in L_2(Y)$, call $S^{(N)}$ L_2-consistent if $E[S_N^{(N)}(X)-E(\phi(Y)|X)]^2 \to 0$. To see that our definition is equivalent, put down the fixed point (x,y) and then the other $N-1$ random points $(x_1,y_1),\ldots,(x_{N-1},y_{N-1})$. Now compute $E[S_N^{(N)}(x)-E(Y|x)]^2 = g_N(x)$. Our definition of m.s. consistency is then

$$E[\tfrac{1}{n} \textstyle\sum_k g_N(x_k)] \to 0$$

or $Eg_N(X) \to 0$.

Uniform boundedness is a critical condition for consistency proofs. A key element in Stone's proof is (put in different form)

(A.1) PROPOSITION. *Take data sets drawn from a bivariate distribution* (X,Y), $S^{(N)}$ *a uniformly bounded sequence of smooths on* x, P_X *the conditional expectation operator. If, for a set of functions* $\{\phi\}$ *dense in* $H_2(Y)$,

$$E\|S^{(N)}(\phi|x)-P_X(\phi|x)\|_N^2 \to 0 ,$$

then the $S^{(N)}$ *as m.s. consistent. If, for a set of functions* $\{h\}$ *dense in* $H_2(X)$,

(A.2) $$E\|S^{(N)}(h|x)-h(x)\|_N^2 \to 0$$

then (A.2) holds for all $h \in H_2(X)$.

The proof is simple and is omitted.

Assume $S^{(N)}$ is linear. Then

(A.3) $$E\|S^{(N)}\phi-P_X\phi\|_N^2 \le 2E\|S^{(N)}(\phi-P_X\phi)\|_N^2 + 2E\|S^{(N)}P_X\phi-P_X\phi\|_N^2 .$$

If it can be shown that $E\|S^{(N)}h-h\|_N^2 \to 0$ for all continuous $h \in H_2(X)$ vanishing off at finite intervals, and if the first term on the right in (A.3) goes to zero for all ϕ such that $\|\phi\|_\infty < \infty$, then (A.1) implies that $S^{(N)}$ is m.s. consistent. This strategy works for a wide variety of smooths.

To illustrate, because Stone's results [1977] do not seem immediately applicable to bivariate regression smooths, m.s. consistency is proven for a modified regression smooth similar to supersmoother. For x any point, let $J(x)$ be the indices of the M points in $\{x_k\}$ directly above x plus the M below. If there are only $M' < M$ above (below) then include the $M+(M-M')$ directly below (above). For a regression smooth

(A.4) $$S(\phi|x) = \bar{\phi}_x + \frac{\Gamma_x(\phi,x)}{\sigma_x^2}(x-\bar{x}_x)$$

where $\bar{\phi}_x$, $\bar{x}_x$ are the averages of $\phi(y_k)$, x_k over the indices in $J(x)$,

$\Gamma_x(\phi,x)$, σ_x^2 the covariance between $\phi(y_k)$, x_k and the variance of x_k over the indices in $J(x)$.

Write the second term in (A.3) as

$$\frac{\Gamma_x(\phi,x)}{\sigma_x}\,\frac{(x-\bar{x}_x)}{\sigma_x}\ .$$

If there are M points above and below in $J(x)$, it is not hard to show that

$$\left|\frac{x-\bar{x}_x}{\sigma_x}\right| \le 1.$$

This is not true near endpoints where $(x-\bar{x}_x)/\sigma_x$ can become arbitrarily large as M gets large. This endpoint behavior keeps regression from being uniformly bounded. To remedy this, define a function

$$[x]_t = \begin{cases} x\ , & |x| \le 1 \\ \mathrm{sgn}(s)\ , & |x| > 1\ , \end{cases}$$

and define the *modified regression smooth* by

$$S(\phi|x) = \bar{\phi}_x + \frac{\Gamma_x(\phi,x)}{\sigma_x}\left[\frac{x-\bar{x}_x}{\sigma_x}\right]_t \qquad \text{(A.5)}$$

This modified smooth is bounded by 2.

(A.6) THEOREM. *If, as* $N\to\infty$, $M\to\infty$, $M \log N/N \to 0$ *and* $P(dx)$ *has no atoms, then the modified regression smooths are m.s. consistent.*

PROOF. Assume $\|\phi\|_\infty < \infty$ and use the inequality (A.3) with $g(x) = P_x(\phi|x)$. Then

$$S(\phi-g|x) = \frac{1}{2M}\sum_{j\in I(x)}(\phi(y_j)-g(x_j))\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}\left[\frac{x-\bar{x}_x}{\sigma_x}\right]_t\}\ .$$

The conditional expectation of $[S(\phi-g|x)]^2$ given $\{x_k\}$ is

$$\frac{1}{4M^2}\sum_{j\in I(x)}E[(\phi(y_j)-g(x_j))^2|x_j]\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}[\frac{x-\bar{x}_x}{\sigma_x}]\}^2$$
$$\leq \frac{1}{M}\|\phi\|_\infty .$$

Thus, the first term in (A.3) is asymptotically zero. Now look at $S(h|x)-h(x)$, $h\in H_2(X)$, h continuous and zero outside a finite interval;

$$S(h|x)-h(x)=\frac{1}{2M}\sum_{j\in J(x)}(h(x_j)-h(x))\{1+\frac{(x_j-\bar{x}_x)}{\sigma_x}[\frac{x-\bar{x}_x}{\sigma_x}]_t\}$$

Then, for $H(\delta)=\max\{|h(x')-h(x")|;\ |x'-x"|\leq\delta\}$,

$$[S(h|x)-h(x)]^2\leq[\frac{1}{2M}\sum_{j\in J(x)}(h(x_j)-h(x))^2]\cdot 2$$
$$\leq 2H(\max_{j\in J(x)}|x_j-x|) .$$

Then to get $E[S(h|x)-h(x)]^2\to 0$, it is enough to show that $\Delta_N=\max\{|x_j-x|;\ x_j\in J(x)\}$ converges in probability to zero. Take x to be a point such that $P[(x,x+\varepsilon)]>0$, $P[(x-\varepsilon,x)]>0$ for all $\varepsilon>0$. The set S of all such points has $P(S)=1$. Then

$$\{\Delta_N>\varepsilon\}\subset\{\text{at most } 2M-1 \text{ of } \{x_k\} \text{ in } (x-\varepsilon,x)\}$$
$$\cup\{\text{at most } 2M-1 \text{ of } \{x_k\} \text{ in } (x,x+\varepsilon)\} .$$

Using the Binomial distribution gives the bound

$$P(\Delta_N>\varepsilon)\leq 2N^M\{(1-P[(x,x+\varepsilon)])^{N-M}+(1-P[(x-\varepsilon,x)])^{N-M}\}$$

Holding ε fixed with $N\to\infty$ and $M=o(N/\log N)$ results in $P(\Delta_N>\varepsilon)\to 0$, proving the theorem.

References

Anscombe, F.J. and Tukey, J.W. (1963). The examination and analysis of residuals. *Technometrics* 5, 141-160.

Belsey, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*, John Wiley and Sons.

Box, G.E.P. and Tidwell, P.W. (1962). Transformations of the independent variables. *Technometrics* 4, 531-550.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J.R. Statist. Soc. B* 26, 211-252.

Box, G.E.P. and Hill, W.J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics* 16, 385-389.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74, 828-836.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31, 317-403.

deBoor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* 9, 1310-1319.

Devroye, L. and Wagner, T.J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* 8, 231-239.

Draper, N.R. and Cox, D.R. (1969). On distributions and their transformations to normality. *J.R. Statist. Soc. B* 31, 472-476.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1-26.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Ann. Math. Statist.* 38, 1456-1465.

Friedman, J.H. and Stuetzle, W. (1982). Smoothing of scatterplots. Dept. of Statistics, Stanford University, Tech. Report ORION006.

Gasser, T. and Rosenblatt, M. (eds.) (1979). *Smoothing Techniques for Curve Estimation*, in Lecture Notes in Mathematics 757, New York: Springer-Verlag.

Gebelein, H. (1941). Das statitistiche problem der korrelation als variations und eigenwert problem und sein Zusammenhang mit der Ausgleichung-srechnung. *Z. Angew. Math. Mech.* 21, 364-379.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Mngmnt* 5, 81-102.

Kendall, M.A. and Stuart, A. (1967). *The Advanced Theory of Statistics*, Volume 2, Hafner.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115-129.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *J.R. Statist. Soc. B* 27, 251-263.

Linsey, J.K. (1972). Fitting response surfaces with power transformations. *J.R. Statist. Soc. C* 21, 234-237.

Linsey, J.K. (1974). Construction and comparison of statistical models. *J.R. Statist. Soc. B* 36, 418-425.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression*, Addison-Wesley.

Renyi, A. (1959). On measures of dependence. *Acta. Math. Acad. Sci. Hungar.* 10, 441-451.

Sarmanov, O.V. (1958a). The maximal correlation coefficient (symmetric case). *Dokl. Acad. Nauk. SSSR* 120, 715-718.

Sarmanov, O.V. (1958b). The maximal correlation coefficient (nonsymmetric case). *Dokl. Acad. Nauk. SSSR* 121, 52-55.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* 8, 240-246.

Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* 7, 139-149.

Tukey, J.W. (1982). The use of smelting in guiding re-expression, in *Modern Data Analysis*, Laurner and Siegel (eds.), Academic Press.

Wood, J.T. (1974). An extension of the analysis of transformations of Box and Cox. *Appl. Statist. (J.R. Statist. Soc. C)* 23.

TABLE 1

Variables Used in the Housing Value Equation
of Harrison and Rubinfeld (1978)

Variable	Definition
MV	Median value of owner-occupied homes
RM	Average number of rooms in owner units
AGE	Proportion of owner units built prior to 1940
DIS	Weighted distances to five employment centers in the Boston region
RAD	Index of accessibility to radial highways
TAX	Full property tax rate ($/$10,000)
PTRATIO	Pupil-teacher ratio by town school district
B	Black proportion of population
LSAT	Proportion of population that is lower status
CRIM	Crime rate by town
ZN	Proportion of town's residential land zoned for lots greater than 25,000 square feet
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy: = 1 if tract bounds the Charles River; = 0 if otherwise
NOX	Nitrogen oxide concentration in pphm

Appendix 2

FORTRAN Implementation of ACE Algorithm

This section presents a listing of a FORTRAN program implementing the ACE algorithm. It consists of six subroutines. The first two are called by the user to perform functions (ACE,ACEMOD), the next is a BLOCK DATA subprogram in which the values of several internal parameters are initialized, and the last three provide utilities used by the first two subroutines. The user interface is through FORTRAN subroutine calls. The data and various input parameters are supplied as arguments in the calling sequence. The optimal transformation estimates and other output quantities, as well as required scratch storage are also passed as subroutine arguments. In order to employ these routines it is necessary to write a main or calling program that reads the input data into appropriate arrays, to declare additional arrays for output and scratch storage, and then to execute a call to the appropriate subroutine (ACE or ACEMOD).

SUBROUTINE ACE computes the optimal transformation estimates using the double loop restart version of the ACE algorithm (see Sections 2 and 5.3). These estimates are stored as a set of transformed values, one value for each observation, for the response and each predictor variable. Upon return from SUBROUTINE ACE these values are stored in the arrays specified in the subroutine call. This information can then be passed to appropriate graphics routines for display, or to function approximation routines for summarization. SUBROUTINE ACEMOD can (optionally) be called to estimate new response values given a set of predictor values and the optimal transformation estimates from SUBROUTINE ACE.

As written, this program is intended to be used in conjunction with a particular smoothing subroutine ("super-smoother") which is listed in Friedman and Stuetzle (1982). Our experience so far with the ACE procedure has been gained in this context. It is possible to employ other smoothing routines by properly interfacing them with the ACE code. ACE calls for the smooth by the FORTRAN statement

```
CALL SUPSMU (N,X,Y,W,L,ALPHA,RESPAN,IBIN,SMO,SCR)
```

where the parameters have the following meaning:

N: number of observations (X,Y) pairs.

X(1...N): ordered abscissa values.

Y(1...N): corresponding ordinate values.

W(1...N): weight for each observation.

L: abscissa variable flag - L = 1 ordered variable.
L = 2 periodic (circular) variable.

ALPHA,RESPAN,IBIN: miscellaneous parameters (can be set through COMMON/PARMS/).

SMO(1...N): output smoothed ordinate values.

SCR(1...N,1...3): scratch array.

```
      SUBROUTINE ACE (P,N,X,Y,W,L,DELRSQ,TX,TY,RSQ,IERR,M,Z)
C-----------------------------------------------------------------------
C
C OPTIMAL TRANSFORMATIONS FOR CORRELATION AND MULTIPLE REGRESSION
C BY ALTERNATING CONDITIONAL ESTIMATES.
C
C (BREIMAN AND FRIEDMAN, 1982).
C
C CODED BY: J. H. FRIEDMAN
C           DEPARTMENT OF STATISTICS AND
C           STANFORD LINEAR ACCELERATOR CENTER
C           STANFORD UNIVERSITY
C           STANFORD, CA. 94305
C
C INPUT:
C
C    N : NUMBER OF OBSERVATIONS.
C    P : NUMBER OF PREDICTOR VARIABLES FOR EACH OBSERVATION.
C    X(P,N) : PREDICTOR DATA MATRIX.
C    Y(N) : RESPONSE VALUE FOR EACH OBSERVATION.
C    W(N) : WEIGHT FOR EACH OBSERVATION.
C    L(P+1) : FLAG FOR EACH VARIABLE.
C       L(1) THROUGH L(P) : PREDICTOR VARIABLES.
C       L(P+1) : RESPONSE VARIABLE.
C       L(I)=0 => ITH VARIABLE NOT TO BE USED.
C       L(I)=1 => ITH VARIABLE ASSUMES ORDERED VALUES.
C       L(I)=2 => ITH VARIABLE ASSUMES CIRCULAR (PERIODIC) VALUES.
C       L(I)=3 => ITH VARIABLE TRANSFORMATION IS TO BE MONOTONE.
C       L(I)=4 => ITH VARIABLE TRANSFORMATION IS TO BE LINEAR.
C       L(I)=5 => ITH VARIABLE ASSUMES CATEGORICAL VALUES.
C   DELRSQ : TERMINATION THRESHOLD. ITERATION STOPS WHEN
C       RSQ CHANGES LESS THAN DELRSQ IN NTERM
C       CONSECUTIVE ITERATIONS (SEE BELOW - DEFAULT, NTERM=3).
C
C OUTPUT:
C
C   TX(N,P) : PREDICTOR TRANSFORMATIONS.
C      TX(J,I) = TRANSFORMED VALUE OF ITH PREDICTOR FOR JTH OBS.
C   TY(N) = RESPONSE TRANSFORMATION.
C      TY(J) = TRANSFORMED RESPONSE VALUE FOR JTH OBSERVATION.
C   RSQ = FRACTION OF VARIANCE(TY<Y>)
C                          P
C         EXPLAINED BY SUM TX(I)<X(I)> .
C                        I=1
C   IERR : ERROR FLAG.
C      IERR = 0 : NO ERRORS DETECTED.
C      IERR > 0 : ERROR DETECTED - SEE FORMAT STATEMENTS BELOW.
C
C SCRATCH:
C
C    M(N,P+1), Z(N,7) : INTERNAL WORKING STORAGE.
C    (Z(J,1), J=1,N) CONTAIN (TRANSFORMED) RESIDUALS AS OUTPUT.
C
C NOTE: THIS ROUTINE USES AS A PRIMITIVE THE 'SUPER SMOOTHER'
C    SUPSMU (SEE - FRIEDMAN AND STUETZLE (1982). SMOOTHING OF
```

```
C     SCATTERPLOTS. STANFORD UNIVERSITY STATISTICS DEPARTMENT
C     REPORT ORION006.)
C
C-----------------------------------------------------------------
      INTEGER P,PP1,M(N,1),L(1)
      REAL Y(N),X(P,N),W(N),TY(N),TX(N,P),Z(N,7),CT(10)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,ALPHA,RESPAN,IBIN
      DOUBLE PRECISION SM,SV,SW
      IERR=0
      PP1=P+1
      SM=0.0
      SV=SM
      SW=SV
      IF (L(PP1).GT.0) GO TO 10
      IERR=4
      IF (ITAPE.GT.0) WRITE (ITAPE,360) PP1
      RETURN
 10   NP=0
      DO 20 I=1,P
      IF (L(I).GT.0) NP=NP+1
 20   CONTINUE
      IF (NP.GT.0) GO TO 30
      IERR=5
      IF (ITAPE.GT.0) WRITE (ITAPE,370) P
      RETURN
 30   DO 40 J=1,N
      SM=SM+W(J)*Y(J)
      SV=SV+W(J)*Y(J)**2
      SW=SW+W(J)
      M(J,PP1)=J
      Z(J,2)=Y(J)
 40   CONTINUE
      IF (SW.GT.0.0) GO TO 50
      IERR=1
      IF (ITAPE.GT.0) WRITE (ITAPE,330)
      RETURN
 50   SM=SM/SW
      SV=SV/SW-SM**2
      IF (SV.LE.0.0) GO TO 60
      SV=1.0/DSQRT(SV)
      GO TO 70
 60   IERR=2
      IF (ITAPE.GT.0) WRITE (ITAPE,340)
      RETURN
 70   DO 80 J=1,N
      Z(J,1)=(Y(J)-SM)*SV
 80   CONTINUE
      CALL SORT (Z(1,2),M(1,PP1),1,N)
      DO 100 I=1,P
      IF (L(I).LE.0) GO TO 100
      DO 90 J=1,N
      TX(J,I)=0.0
      M(J,I)=J
      Z(J,2)=X(I,J)
 90   CONTINUE
```

```
      CALL SORT (Z(1,2),M(1,I),1,N)
  100 CONTINUE
      RSQ=0.0
      ITER=0
      NTERM=MIN0(NTERM,10)
      NT=0
      DO 110 I=1,NTERM
      CT(I)=100.0
  110 CONTINUE
  120 ITER=ITER+1
      DO 130 J=1,N
      TY(J)=Z(J,1)
  130 CONTINUE
      NIT=0
  140 RSQI=RSQ
      NIT=NIT+1
      DO 200 I=1,P
      IF (L(I).LE.0) GO TO 200
      DO 150 J=1,N
      K=M(J,I)
      Z(J,1)=TY(K)+TX(K,I)
      Z(J,2)=X(I,K)
      Z(J,7)=W(K)
  150 CONTINUE
      CALL SMOTHR (L(I),N,Z(1,2),Z,Z(1,7),Z(1,6),Z(1,3))
      SM=0.0
      DO 160 J=1,N
      SM=SM+Z(J,7)*Z(J,6)
  160 CONTINUE
      SM=SM/SW
      DO 170 J=1,N
      Z(J,6)=Z(J,6)-SM
  170 CONTINUE
      SV=0.0
      DO 180 J=1,N
      SV=SV+Z(J,7)*(Z(J,1)-Z(J,6))**2
  180 CONTINUE
      SV=1.0-SV/SW
      IF (SV.LE.RSQ) GO TO 200
      RSQ=SV
      DO 190 J=1,N
      K=M(J,I)
      TX(K,I)=Z(J,6)
      TY(K)=Z(J,1)-Z(J,6)
  190 CONTINUE
  200 CONTINUE
      IF ((NP.NE.1).AND.((RSQ-RSQI.GT.DELRSQ).AND.(NIT.LT.MAXIT))) GO TC.
     1 140
      IF (RSQ.NE.0.0.OR.ITER.NE.1) GO TO 230
      DO 220 I=1,P
      IF (L(I).LE.0) GO TO 220
      DO 210 J=1,N
      TX(J,I)=X(I,J)
  210 CONTINUE
  220 CONTINUE
```

```
230   DO 250 J=1,N
      K=M(J,PP1)
      Z(J,2)=Y(K)
      Z(J,7)=W(K)
      Z(J,1)=0.0
      DO 240 I=1,P
      IF (L(I).GT.0) Z(J,1)=Z(J,1)+TX(K,I)
240   CONTINUE
250   CONTINUE
      CALL SMOTHR (L(PP1),N,Z(1,2),Z,Z(1,7),Z(1,6),Z(1,3))
      SM=0.0
      SV=SM
      DO 260 J=1,N
      K=M(J,PP1)
      SM=SM+W(K)*Z(J,6)
      SV=SV+W(K)*Z(J,6)**2
      Z(K,2)=Z(J,1)
260   CONTINUE
      SM=SM/SW
      SV=SV/SW-SM**2
      IF (SV.LE.0.0) GO TO 270
      SV=1.0/DSQRT(SV)
      GO TO 280
270   IERR=3
      IF (ITAPE.GT.0) WRITE (ITAPE,350)
      RETURN
280   DO 290 J=1,N
      K=M(J,PP1)
      TY(K)=(Z(J,6)-SM)*SV
290   CONTINUE
      SV=0.0
      DO 300 J=1,N
      Z(J,1)=TY(J)-Z(J,2)
      SV=SV+W(J)*Z(J,1)**2
300   CONTINUE
      RSQ=1.0-SV/SW
      IF (ITAPE.GT.0) WRITE (ITAPE,320) ITER,RSQ
      NT=MOD(NT,NTERM)+1
      CT(NT)=RSQ
      CMN=100.0
      CMX=-100.0
      DO 310 I=1,NTERM
      CMN=AMIN1(CMN,CT(I))
      CMX=AMAX1(CMX,CT(I))
310   CONTINUE
      IF ((CMX-CMN.GT.DELRSQ).AND.(ITER.LT.MAXIT)) GO TO 120
      RETURN
320   FORMAT(    11H ITERATION I2,    23H    R**2  =  1 - E**2  =G12.4)
330   FORMAT(    41H IERR=1: SUM OF WEIGHTS (W) NOT POSITIVE.)
340   FORMAT(    29H IERR=2: Y HAS ZERO VARIANCE.)
350   FORMAT(    30H IERR=3: TY HAS ZERO VARIANCE.)
360   FORMAT(    11H IERR=4: L(I2,    19H) MUST BE POSITIVE.)
370   FORMAT(    29H IERR=5: AT LEAST ONE L(1)-L(I2,    19H) MUST BE POSI'
     1IVE.)
      END
```

```
      SUBROUTINE ACEMOD (V,P,N,X,Y,L,TX,TY,M,YHAT,IERR)
C-----------------------------------------------------------------------
C
C COMPUTES RESPONSE ESTIMATES FROM THE MODEL
C                          -1    P
C                 YHAT = TY   ( SUM TX(I)<V(I)> )
C                              I=1
C USING THE TRANSFORMATIONS CONSTRUCTED BY SUBROUTINE ACE.
C
C INPUT:
C
C    V(P) : VECTOR OF PREDICTOR VALUES.
C    P,N,X,Y,L : SAME INPUT AS FOR SUBROUTINE ACE.
C    TX,TY,M : OUTPUT FROM SUBROUTINE ACE.
C
C OUTPUT:
C
C   YHAT : ESTIMATED RESPONSE VALUE FOR V.
C   IERR : ERROR FLAG.
C      IERR=0: NO ERROR DETECTED.
C      IERR=1: ERROR DETECTED - SEE FORMAT STATEMENT BELOW.
C
C NOTE: THIS ROUTINE REQUIRES THAT THE RESPONSE TRANSFORMATION TY IS A
C    STRICTLY (INCREASING OR DECREASING) MONOTONE FUNCTION OF Y, THAT
C    IS L(P+1) = 3 OR 4 IN THE CALL TO SUBROUTINE ACE.
C
C-----------------------------------------------------------------------
      INTEGER P,PP1,M(N,1),L(1),LOW,HIGH,PLACE
      REAL V(P),X(P,N),Y(N),TY(N),TX(N,P)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,ALPHA,RESPAN,IBIN
      PP1=P+1
      IERR=0
      IF (L(PP1).EQ.3.OR.L(PP1).EQ.4) GO TO 10
      IERR=1
      IF (ITAPE.GT.0) WRITE (ITAPE,140) PP1
      RETURN
 10   YH=0.0
      DO 80 I=1,P
      IF (L(I).LE.0) GO TO 80
      VI=V(I)
      IF (VI.GT.X(I,M(1,I))) GO TO 20
      PLACE=1
      GO TO 70
 20   IF (VI.LT.X(I,M(N,I))) GO TO 30
      PLACE=N
      GO TO 70
 30   LOW=0
      HIGH=N+1
 40   IF (LOW+1.GE.HIGH) GO TO 60
      PLACE=(LOW+HIGH)/2
      XT=X(I,M(PLACE,I))
      IF (VI.EQ.XT) GO TO 70
      IF (VI.GE.XT) GO TO 50
      HIGH=PLACE
      GO TO 40
```

```
 50   LOW=PLACE
      GO TO 40
 60   JL=M(LOW,I)
      JH=M(HIGH,I)
      YH=YH+TX(JL,I)+(TX(JH,I)-TX(JL,I))*(VI-X(I,JL))/(X(I,JH)-X(I,JL))
      GO TO 80
 70   YH=YH+TX(M(PLACE,I),I)
 80   CONTINUE
      IF (YH.GT.TY(M(1,PP1))) GO TO 90
      YHAT=Y(M(1,PP1))
      RETURN
 90   IF (YH.LT.TY(M(N,PP1))) GO TO 100
      YHAT=Y(M(N,PP1))
      RETURN
 100  LOW=0
      HIGH=N+1
      XT=TY(M(N,PP1))-TY(M(1,PP1))
      XT=XT/ABS(XT)
 110  IF (LOW+1.GE.HIGH) GO TO 130
      PLACE=(LOW+HIGH)/2
      IF (XT*YH.GE.XT*TY(M(PLACE,PP1))) GO TO 120
      HIGH=PLACE
      GO TO 110
 120  LOW=PLACE
      GO TO 110
 130  JL=M(LOW,PP1)
      JH=M(HIGH,PP1)
      YHAT=Y(JL)+(Y(JH)-Y(JL))*(YH-TY(JL))/(TY(JH)-TY(JL))
      RETURN
 140  FORMAT(    11H IERR=1: L(I2,      55H) MUST EQUAL 3 OR 4 - MONOTONE
     1RESPONSE TRANSFORMATION.)
      END
      BLOCK DATA
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,ALPHA,RESPAN,IBIN
C-----------------------------------------------------------------------
C
C THESE PROCEDURE PARAMETERS CAN BE CHANGED IN THE CALLING ROUTINE
C BY DEFINING THE ABOVE LABELED COMMON AND RESETTING THE VALUES WITH
C EXECUTABLE STATEMENTS.
C
C ITAPE : FORTRAN FILE NUMBER FOR PRINTER OUTPUT.
C         (ITAPE.LE.0 => NO PRINTER OUTPUT.)
C MAXIT : MAXIMUM NUMBER OF ITERATIONS.
C NTERM : NUMBER OF CONSECUTIVE ITERATIONS FOR WHICH ESTIMATED
C         CORRELATION MUST CHANGE LESS THAN DELCOR FOR CONVERGENCE.
C ALPHA, RESPAN, IBIN : SUPER SMOOTHER PARAMETERS.
C   (SEE - FRIEDMAN AND STUETZLE, REFERENCE ABOVE.)
C
C-----------------------------------------------------------------------
      DATA ITAPE,MAXIT,NTERM,ALPHA,RESPAN,IBIN /6,20,3,0.1,0.25,1/
      END
      SUBROUTINE SMOTHR (L,N,X,Y,W,SMO,SCR)
      REAL X(N),Y(N),W(N),SMO(N),SCR(N,3)
      COMMON /PARMS/ ITAPE,MAXIT,NTERM,ALPHA,RESPAN,IBIN
      DOUBLE PRECISION SM,SW
```

```
      IF (L.LT.5) GO TO 50
      J=1
10    J0=J
      SM=W(J)*Y(J)
      SW=W(J)
      IF (J.GE.N) GO TO 30
20    IF (X(J+1).GT.X(J)) GO TO 30
      J=J+1
      SM=SM+W(J)*Y(J)
      SW=SW+W(J)
      IF (J.LT.N) GO TO 20
30    SM=SM/SW
      DO 40 I=J0,J
      SMO(I)=SM
40    CONTINUE
      J=J+1
      IF (J.LE.N) GO TO 10
      GO TO 240
50    IF (L.NE.4) GO TO 80
      SM=0.0
      SW=SM
      DO 60 J=1,N
      SM=SM+W(J)*X(J)*Y(J)
      SW=SW+W(J)*X(J)**2
60    CONTINUE
      A=SM/SW
      DO 70 J=1,N
      SMO(J)=A*X(J)
70    CONTINUE
      GO TO 240
80    CALL SUPSMU (N,X,Y,W,L,ALPHA,RESPAN,IBIN,SMO,SCR)
      IF (L.NE.3) GO TO 240
      DO 90 J=1,N
      SCR(J,1)=SMO(J)
      SCR(N-J+1,2)=SCR(J,1)
90    CONTINUE
      CALL MONTNE (SCR,N)
      CALL MONTNE (SCR(1,2),N)
      SM=0.0
      SW=SM
      DO 100 J=1,N
      SM=SM+(SMO(J)-SCR(J,1))**2
      SW=SW+(SMO(J)-SCR(N-J+1,2))**2
100   CONTINUE
      IF (SM.GE.SW) GO TO 120
      DO 110 J=1,N
      SMO(J)=SCR(J,1)
110   CONTINUE
      GO TO 140
120   DO 130 J=1,N
      SMO(J)=SCR(N-J+1,2)
130   CONTINUE
140   J=1
150   J0=J
      IF (J.GE.N) GO TO 170
```

```
160   IF (SMO(J+1).NE.SMO(J)) GO TO 170
      J=J+1
      IF (J.LT.N) GO TO 160
170   IF (J.LE.J0) GO TO 190
      A=0.0
      IF (J0.GT.1) A=0.5*(SMO(J0)-SMO(J0-1))
      B=0.0
      IF (J.LT.N) B=0.5*(SMO(J+1)-SMO(J))
      D=(A+B)/(J-J0)
      IF (A.EQ.0.0.OR.B.EQ.0.0) D=2.0*D
      IF (A.EQ.0.0) A=B
      DO 180 I=J0,J
      SMO(I)=SMO(I)-A+D*(I-J0)
180   CONTINUE
190   J=J+1
      IF (J.LE.N) GO TO 150
      J=1
200   J0=J
      SM=SMO(J)
      IF (J.GE.N) GO TO 220
210   IF (X(J+1).GT.X(J)) GO TO 220
      J=J+1
      SM=SM+SMO(J)
      IF (J.LT.N) GO TO 210
220   SM=SM/(J-J0+1)
      DO 230 I=J0,J
      SMO(I)=SM
230   CONTINUE
      J=J+1
      IF (J.LE.N) GO TO 200
240   RETURN
      END
      SUBROUTINE MONTNE (X,N)
      REAL X(N)
      INTEGER BB,EB,BR,ER,BL,EL
      BB=0
      EB=BB
10    IF (EB.GE.N) GO TO 110
      BB=EB+1
      EB=BB
20    IF (EB.GE.N) GO TO 30
      IF (X(BB).NE.X(EB+1)) GO TO 30
      EB=EB+1
      GO TO 20
30    IF (EB.GE.N) GO TO 70
      IF (X(EB).LE.X(EB+1)) GO TO 70
      BR=EB+1
      ER=BR
40    IF (ER.GE.N) GO TO 50
      IF (X(ER+1).NE.X(BR)) GO TO 50
      ER=ER+1
      GO TO 40
50    PMN=(X(BB)*(EB-BB+1)+X(BR)*(ER-BR+1))/(ER-BB+1)
      EB=ER
      DO 60 I=BB,EB
```

```
      X(I)=PMN
60    CONTINUE
70    IF (BB.LE.1) GO TO 10
      IF (X(BB-1).LE.X(BB)) GO TO 10
      BL=BB-1
      EL=BL
80    IF (BL.LE.1) GO TO 90
      IF (X(BL-1).NE.X(EL)) GO TO 90
      BL=BL-1
      GO TO 80
90    PMN=(X(BB)*(EB-BB+1)+X(BL)*(EL-BL+1))/(EB-BL+1)
      BB=BL
      DO 100 I=BB,EB
      X(I)=PMN
100   CONTINUE
      GO TO 30
110   RETURN
      END
```

```
      SUBROUTINE SORT (V,A,II,JJ)
C
C     PUTS INTO A THE PERMUTATION VECTOR WHICH SORTS V INTO
C     INCREASING ORDER.  ONLY ELEMENTS FROM II TO JJ ARE CONSIDERED.
C     ARRAYS IU(K) AND IL(K) PERMIT SORTING UP TO 2**(K+1)-1 ELEMENTS
C
C     THIS IS A MODIFICATION OF CACM ALGORITHM #347 BY R. C. SINGLETON,
C     WHICH IS A MODIFIED HOARE QUICKSORT.
C
      DIMENSION A(JJ),V(1),IU(20),IL(20)
      INTEGER T,TT
      INTEGER A
      REAL V
      M=1
      I=II
      J=JJ
 10   IF (I.GE.J) GO TO 80
 20   K=I
      IJ=(J+I)/2
      T=A(IJ)
      VT=V(IJ)
      IF (V(I).LE.VT) GO TO 30
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      V(IJ)=V(I)
      V(I)=VT
      VT=V(IJ)
 30   L=J
      IF (V(J).GE.VT) GO TO 50
      A(IJ)=A(J)
      A(J)=T
      T=A(IJ)
      V(IJ)=V(J)
      V(J)=VT
      VT=V(IJ)
      IF (V(I).LE.VT) GO TO 50
      A(IJ)=A(I)
      A(I)=T
      T=A(IJ)
      V(IJ)=V(I)
      V(I)=VT
      VT=V(IJ)
      GO TO 50
 40   A(L)=A(K)
      A(K)=TT
      V(L)=V(K)
      V(K)=VTT
 50   L=L-1
      IF (V(L).GT.VT) GO TO 50
      TT=A(L)
      VTT=V(L)
 60   K=K+1
      IF (V(K).LT.VT) GO TO 60
      IF (K.LE.L) GO TO 40
```

```
      IF (L-I.LE.J-K) GO TO 70
      IL(M)=I
      IU(M)=L
      I=K
      M=M+1
      GO TO 90
70    IL(M)=K
      IU(M)=J
      J=L
      M=M+1
      GO TO 90
80    M=M-1
      IF (M.EQ.0) RETURN
      I=IL(M)
      J=IU(M)
90    IF (J-I.GT.10) GO TO 20
      IF (I.EQ.II) GO TO 10
      I=I-1
100   I=I+1
      IF (I.EQ.J) GO TO 80
      T=A(I+1)
      VT=V(I+1)
      IF (V(I).LE.VT) GO TO 100
      K=I
110   A(K+1)=A(K)
      V(K+1)=V(K)
      K=K-1
      IF (VT.LT.V(K)) GO TO 110
      A(K+1)=T
      V(K+1)=VT
      GO TO 100
      END
```